The American Economic Review

ARTICLES

MARCH 1981

THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members six times a year, in March, May, June, September, and semimonthly in December.

Dues for 1981, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:
$30.00 for regular members with rank of assistant professor or lower, or with annual income of $14,400, or less;
$36.00 for regular members with rank of associate professor, or with annual income of $14,400 to $24,000;
$42.00 for regular members with rank of full professor, or with annual income above $24,000;
$15.00 for junior members (registered students). Certification must be submitted yearly.
Subscriptions (libraries, institutions, or firms) are $100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.
In countries other than the United States, add $5.00 to cover extra postage.

●Correspondance relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

P17825

THE AMERICAN ECONOMIC REV

ROBERT W. CLOWER
Managing Editor

LOIS J. BAGLEY
Assistant Editor

WILMA ST. JOHN
Production Editor

● Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

● No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

March 1981

VOLUME 71, NUMBER 1

Articles

Shorter Papers

Moses Abramovitz

Welfare Quandaries and Productivity Concerns

By MOSES ABRAMOVITZ*

The early debates over the role of government in economic life, at least during the era of industrialization, took the form of a contest between *laissez-faire* and thoroughgoing socialism. In Western Europe and North America, however, the movement away from individualism followed a much less radical course, which John Maynard Keynes was one of the first to define. His famous lectures in the mid-1920's on *The End of Laissez-Faire* carried the following passage:

> ...a time may be coming when we shall get clearer than we are at present as to when we are talking about Capitalism as an efficient or inefficient technique, and when we are talking about it as desirable or objectionable in itself. For my part, I think that Capitalism, *wisely managed*, can probably be made more efficient for attaining economic ends than any alternative yet in sight, but that in itself it is in many ways extremely objectionable. Our problem is to work out a social organization which shall be as efficient as possible without offending our notions of a satisfactory way of life. [p. 53, emphasis added]

Keynes, as we can now see, was among the first writers to form a definite vision of the kind of system under which we have come to live during the last half century, the system we now call the Mixed Economy or Welfare Capitalism or the Middle Way. Like the much more individualistic, much less guided, system that preceded it, the Mixed Economy developed with the support of a broad consensus of opinion. That consensus, however, has now weakened. The economic role of government is again the subject of debate, attack, and resistance far more intense than we have known for decades. The attack ranges over a wide spectrum. It questions the scope of government, the particular measures and policies through which government exercises its functions, and the political institutions which shape the measures and policies employed. A few voices call on us to move on to a more encompassing socialism, including the ownership of industry. Many more call for a drastic revival of market rule.

We all, I think, sense that we have come to a very difficult juncture in the development of our Mixed Economy. How we shall emerge is still in dim prospect. As in other illnesses, social crises often are surmounted and are followed by periods of renewed stable development. But sometimes not. We, therefore, ought to think where we are and what the nature of our troubles is.

I

There is no single, simple way to gather together all the threads of our present discontent, and I shall not try. One useful opening, however, is to consider the pronounced and worrisome retardation of productivity growth from which we now suffer. Productivity growth, I need hardly say, is the main source of measured per capita output growth. And per capita output, in turn, is a central component of economic welfare as we economists conceive it, many would say *the* central component. It is elementary, however, that per capita output growth and welfare growth are not the same thing. National product is not even an adequate long-term measure of net output relevant to welfare. It makes inadequate al-

*Presidential address delivered at the ninety-third meeting of the American Economic Association, September 6, 1980, Denver, Colorado. I am grateful to a number of friends who read early drafts and suggested improvements: Kenneth Arrow, Michael Boskin, Arthur Burns, Paul David, Solomon Fabricant, Eli Ginzberg, Bert Hickman, Milton Katz, Ronald McKinnon, Nathan Rosenberg, John Shoven, Robert Solow, and Tibor Scitovsky.

lowance for the quality and variety of goods. It excludes the household and treats all government expenditure as final product. It neglects the externalities of production and consumption and the costs of growth proper, for example, the dislocation of people. It makes dubious assumptions about people's ability to appraise and guard against the dangers carried by jobs and products. And there is much more to economic welfare than can be captured in any long-term measure of output: job stability; income security, a fair distribution of opportunities and rewards.

The economic role of government expanded during the last half century and more in large part in order to pursue the social objectives that are not comprehended in measured net national product. The result is the mixed economy or welfare state in which we now live and which is now the object of attack.

Productivity growth is a useful focus of discussion in relation to the current discontents and the accompanying reappraisal of our mixed economy for a combination of two reasons.

To begin with, productivity, viewed as a source of private earnings, exists in a state of uneasy tension with the other welfare objectives, which we pursue largely through the government. The causes of the tension need to be underscored.

First—an obvious point—the more income that is diverted to social uses, the less of any given aggregate remains under the private control of income earners for their own personal use.

Next, the size of the diversion and the way it is made and used affects the level of output and productivity, present and future. That is partly because a host of government activities are supportive of current output and productivity, and many activities, including some, like education, that are undertaken for generalized social objectives, are in the nature of capital formation.[1] In a still more basic sense, moreover, and one much neglected in current debates, the pace of growth in a country depends not only on its access to new technology, but on its ability to make and absorb the social adjustments required to exploit new products and processes. Simply to recall the familiar, the process includes the displacement and redistribution of populations among regions and from farm to city. It demands the abandonment of old industries and occupations, and the qualification of workers for new, more skilled occupations. The extension of education, with all its implications for shifts in social status, in aspiration, and in political power, is a requisite. Along the technological path which we have followed, growth also demands very large-scale enterprise which establishes new types of market power and alters the relations of workers and employers. Viewed from another angle, the dependent employment status of workers and the mobility of industry and people imply a great change in the structure of families and in their roles in caring for children, the sick, and the old. Because the required adaptations can and do alter the positions, prospects, and power of established groups, conflict and resistance are intrinsic to the growth process. To resolve such conflict and resistance in a way which preserves a large consensus for growth, yet does not impose a cost which retards growth unduly, a mechanism of conflict resolution is needed. The national sovereign state necessarily becomes the arbiter of group conflict and the mitigator of those negative effects of economic change which would otherwise induce resistance to growth.[2]

The enlargement of the government's economic role, including its support of income minima, health care, social insurance, and the other elements of the welfare state, was, therefore—at least up to a point—not just a question of compassionate regard for the unfortunate, and not just a question of reducing inequalities of outcome and opportunity, though that is how people usually think of it. It was, and is—up to a point—a

[1] In 1976, government gross capital formation, including investment in human capital, was estimated to be just a trifle *larger*, 2 percent, than conventional gross private domestic investment (see Robert Eisner).

[2] Compare Simon Kuznets.

part of the productivity growth process itself.

And yet, manifestly, there is another side to the story, the side that is so much to the fore today. The government's roles as referee and as mitigator of the costs of growth—as well as instrument for pursuing welfare goals supplementary to measured productivity—must be paid for. But it is essentially impossible to design a tax system that places no marginal burden on the rewards for productive effort, or a regulatory system that has no cost in measured output. Similarly, we can hardly design a transfer system which—up to a point—necessarily divorces income from work, but which yet does not qualify economic incentives. There is a presumption, therefore, that the tax-transfer-regulatory system, whatever its essential, long-term, indirect, supportive role, operates more immediately and directly to constrict work, saving, investment, and mobility—just how much is, of course, a question.

There is, therefore, an uneasy many-faceted tension between measured productivity growth and the private earnings it generates on the one side, and the pursuit of other welfare goals through government on the other side. The tension implies a difficult and delicate problem of choice and balance. A balance—certainly a wide acceptance of the pace and nature of our joint pursuit of different welfare goals—seemed to exist during the first two postwar decades when productivity growth was relatively rapid. That balance, if it was a balance, has, however, now been upset by the protracted retardation of productivity growth during the last dozen or more years. That is the second reason why productivity growth is a useful focus for examining the current dissatisfaction with our mixed economy.

I shall deal briefly with three matters:

1) What were the developments which were antecedent to (which stand in the background of) our present troubles and its accompanying discontent?

2) What can we now say about the causes of the current productivity retardation? In particular, to what extent is the retardation connected with the enlarged role of government and its pursuit of alternative social goals?

3) What is the outlook for productivity growth, and what are the implications of that outlook for the further development of our mixed economic system?

II

In the early part of the postwar period, economic growth, in the aggregate and per capita, established itself as a premier goal of economic policy—co-equal with "full" employment, perhaps of even higher priority. Besides the standard reason, that per capita growth raises average levels of consumption, there were special reasons. Growth was seen as the best way to overcome poverty without the social conflicts accompanying redistribution. It would create a favorable environment in which to open opportunities for blacks and other minorities. It would provide the resources for meeting still other social goals, for example, extended education and health care. Growth was also sought to maintain defense, to compete politically with a fast-growing Soviet Union and to assert continued leadership in our rapidly progressing alliance. Growth would enable us to help not only the poor in our own country, it would permit us to help the still more impoverished people of the less-developed world. Productivity growth was a goal distinguishable from full employment, but it was also seen—not necessarily correctly—as a condition of full employment. Unless we could hold our own in international trade, our foreign accounts would impose demand restraints on policy and make for chronic underemployment.

This growth, so ardently desired, was in fact achieved. For two decades, income per capita grew faster than ever before and output per hour much faster. At the same time, there was a rapid development of government in pursuit of other welfare objectives, and this was also eagerly sought. The Social Security system established in the 1930's was enlarged; education was rapidly extended; science was fostered; there were large programs for hospital building and housing. The proportion of the population living below defined poverty levels was reduced—the joint result of rising average incomes, and extended insurance and welfare provision. Partly because government was bigger, partly because the scope of progressive taxation was wider, partly because of old age and unemployment insurance and other forms of income maintenance, we enjoyed the benefits of a system of "built-in stabilizers." Recessions were milder and growth more steady than they had ever been before in American experience as an industrialized country.

The main point, however, is that in this period, productivity growth paid easily for the pursuit of other welfare goals. Although government grew faster than *GNP*, fast growth of productivity supported fast growth of per capita disposable income, of real spendable earnings of workers, and of average family incomes.[3] Productivity growth was, therefore, the substantial basis on which the consensus of opinion supporting the development of the mixed economy rested.

III

Frank Knight liked to say that progress is not a question of happiness; it is a question of what people are unhappy about. Not surprisingly, therefore, the progress of the first two postwar decades was followed by a certain recoil from growth—a reordering, if not reversal, of priorities. This took several forms:

1) Whereas in the 1950's, measured growth was regarded as the main instrument for overcoming poverty, as the 1960's wore on the view took hold, with much justification, that future growth alone could not deal adequately with the poverty which past growth had left behind. Although technical progress, capital accumulation, and general education would continue to be important in the future, an increasing proportion of the "residual poor" had special handicaps. They had to be helped directly, principally by a fight against discrimination, by special education and training programs, and by new and expanded schemes for social insurance, income support, health care programs, and other transfers in kind. The impulse to fight poverty directly was fed by new information about the size and composition of the remaining poor population, by the indignation of social reformers and, most of all, by rising racial tensions.[4] "We cannot," said the Council of Economic Advisors, "leave the further wearing away of poverty solely to the general progress of the economy" (1964, p. 60).[5]

2) As individual income levels rose, people generally became more sensitive to their immediate surroundings. They found hospital and educational facilities inadequate and the urban physical plant shabby. Yet the demand for improvement had to be met in difficult circumstances which continue to plague and torment local government to this day. The relative price of public, like that of private, services was rising. Higher incomes and automobiles were transporting upwardly mobile families to the suburbs, carrying their tax base with them. The cities, increasingly abandoned to the poor, unable to tap the suburban affluence about them, could barely cope. Congestion on the highways and streets, noise, air and water pollution, all fed by growth itself, swelled, moved to the countryside and everywhere became more objectionable to otherwise more affluent people.

3) People discovered the terrors of technology—products, working conditions, and environmental changes that carried risks. The dangers feared were often invisible, they operated at a distance and cumulated over time, carrying both real and imaginary threats to health and life now and in generations to come. Technological progress, which for decades had been seen as the process by which problems and dangers might be overcome, was now increasingly feared as a major source of our troubles.

These shifts in outlook had two important practical consequences. One was the very

[3] See Appendix Table 2.

[4] See the articles by Michael Harrington and John Kenneth Galbraith reprinted in Burton Weisbrod (pp. 29–42 and 49–56, respectively).

[5] Compare ch. 2 *passim*, *Economic Report of the President* (1964).

rapid expansion of government social welfare and civil rights programs which began in the mid-1960's and which developed and matured in the 1970's. Expenditures for "social welfare," which were 9 percent of *GNP* in 1950 and only 10 percent in 1960, rose to 15 percent in 1970 and to 20 percent in 1977.[6] The other was "explosion" of public regulatory legislation and administration directed to the protection of the environment, and to the safety of workers and consumers.[7] The new legislation became the basis for strong, privately organized campaigns to limit growth and the application of new technology.

IV

The maturing of the Great Society programs in the spheres of welfare and civil rights, and the implementation and expansion of the social regulatory laws, brought our mixed economy to a new stage of development. There was a new distribution of emphasis among the different dimensions of economic welfare, and correspondingly a new distribution of economic power between the private and public spheres. The new development of the mixed economy, however, is now confronted by a changed and less-favorable growth environment.

Looking back, we can now see that a slower rate of productivity growth accompanied the institution and the maturing of the Great Society programs. To what extent the two developments were associated as effect and cause, however, is still an open question. So is a related matter; that is, the responsibility of transient as distinct from durable factors for bringing about the slowdown we observe. It would be wrong to pretend that there are now definite answers to these questions. The factual position, however, deserves description because it bears on the origins of our present discontents.

Beginning in the late 1960's, private-sector productivity growth fell back from the high speed it had reached in the years preceding. The retardation before 1973 was moderate. The new pace approximated that during the somewhat slack later 1950's. After 1973, however, the slowdown became much more serious. The upshot is that average productivity growth for the fourteen years between 1965 and 1979 ran at only one-half the pace of the years from 1948 to 1965; since 1973, it has risen at less than one-fifth that earlier pace.[8] The extent of the slowdown between the two rough halves of the postwar period, before and after 1965—to say nothing of the post-1973 period by itself—may be judged from the fact that the post-1965 productivity slowdown has been more severe than any of the retardations measured across major depressions going back to the 1890's. That includes the retardation from the 1920's to the 1930's.[9] Yet, up to 1979 we had had no major depression.

In my judgment, the productivity retardation, at least since 1973, has been accompanied by a slower rate of improvement in material conditions of well-being. In some respects, and by some measures, there have even been significant declines. It is true that, because the labor force was rising rapidly in relation to population, the growth rate of real disposable income per capita was well maintained—at least if we depend on the deflator for "personal consumption expenditures"; not if we use the *CPI*. As perceived by many people, however, the welfare sig-

[6] See U.S. Social Security Administration. Social welfare expenditures cover social insurance, public assistance, health and medical care, veterans' programs, education, housing and "other." At present, exhaustive expenditures account for nearly half and transfer programs for somewhat more than half of total welfare expenditures. (See Sheldon Danziger, Robert Haveman, and Robert Plotnick, pp. 6–8.) The major reasons for the accelerated growth since 1965 appear to lie in the initiation and expansion of new programs, such as Medicare, and in the generous increase of benefit schedules in old programs like Social Security (see Plotnick, pp. 277–78).

[7] *The Federal Register*, which records new regulations, contained 10,000 pages in 1953, but 65,000 pages in 1977. The federal budget to administer regulatory activities was $5 billion in 1978, having doubled since 1974. Compare Arthur Burns, p. 4.

[8] I depend for these comparisons on the easily accessible Bureau of Labor Statistics figures for "output per hour of all persons" in the private business sector. See *Economic Report of the President* (1980, Table B-37).

[9] See Appendix Table 1.

ńificance of even the more favorable measure is qualified. That is partly because the demographic changes that supported labor-force growth also made for a faster increase of households than of population, so to some degree expenses per head increased with income per head.[10] It is qualified also to the extent that women felt forced to take paid work to offset the slower rise or actual decline of their husbands' real earnings; to the extent that the proportion of persons living in pretransfer poverty has been tending to rise since 1968; to the extent that transfer incomes became a more important part of aggregate disposable income—to the disadvantage of income earners; and to the extent that the rise of noncash compensation reduced worker's discretionary take-home pay. The upshot is that in recent years, the average real cash incomes of workers have, depending on the measure, almost ceased to rise or begun to fall. The same is true of the average real total income of families, supported as that has been by transfer incomes and by the entry of second workers. The presumption is that the real earned income of representative single worker families, still more their cash income, has definitely declined.[11]

The slowdowns in the growth rates of productivity, annual wages, and household incomes are, moreover, not the only disturbing elements in our economic situation. They are accompanied by rapid and volatile inflation which redistributes income and wealth in arbitrary and confusing ways. Taken together, these developments have disappointed peoples' expectations; they have robbed many people of the fruits of earlier work and saving, and made almost everyone unsure or fearful about their future.

These developments stand in the background of the current discontent with the operation of our mixed economy. They have led to a blacklash against the earlier recoil from productivity growth. This blacklash—perhaps justifiably, perhaps not—raises sharply the issue of maintaining a steady balance between the productivity growth that supports the rise of earned incomes and the pursuit of other vitally important social goals.

V

Our attitudes towards that issue would be clearer if we could know to what extent the current productivity retardation is actually due to the workings of our mixed economy or to its past and current attempts to raise social welfare through government actions. Many believe that the welfare and regulatory programs are heavily implicated both in direct ways and because of their arguably plausible connection with the onset and persistence of an erratic and accelerating inflation. There is a concomitant fear that the welfare and regulatory programs may be a serious drag on future productivity growth. Opposition to these programs, is, therefore, rising. True, if future productivity growth is slow for whatever reasons, people will be less willing than they might otherwise be to bear the cost of pursuing alternative welfare goals. But if that pursuit were actually a significant cause of slower growth, the reluctance would be still stronger, as it then should be.

The causes of the current retardation, however, remain cloudy. A portion of the slowdown is, by general agreement, due to a virtual cessation of the shift of workers from small-scale inefficient farming and from self-employment in petty trade to higher productivity occupations in larger-scale urban enterprise. A portion too is assignable to the massive entry of workers—youth and women—since the mid-1960's. Finally, a small part of the retardation is attributable to the diversion of resources to comply with environmental regulation and safety requirements in ways that do not register in measured output, though, of course, they

[10] Manifestly, some of the faster increase of households than of population was the result of changing tastes, rising incomes, and better provision for old people through Social Security. It was, therefore, the way in which people chose to spend income to best advantage. But part of the fast increase of households was due to the appearance of large cohorts of young adults who were reaching an age when the establishment of independent households was normal, and, in that sense, the extra expense of separate households was imposed on them.

[11] See Appendix Table 2.

should. Serious students, however, offer widely different estimates of the contributions of other factors: the quality of schooling, conventional capital services, *R&D*, and the influence of cyclical or other forces affecting intensity of resource use. The impact of higher energy prices on the substitution of labor for capital in the operation of existing energy-using equipment and on the post-1973 slowdown of capital deepening is equally unclear, though possibly very important. Most analyses leave a substantial part of the retardation unconnected with any identified and measured contributory source, and they disagree about the time—whether after 1973 or as early as the latter 1960's—when that unspecified residual retardation made its appearance.[12]

In this state of factual uncertainty, it is not hard to propose estimates of the sources of retardation which assign substantial responsibility to factors connected with the government's welfare and regulatory activities. We, therefore, find William Fellner asking: "...whether, directly or indirectly [the analyses of the retardation] do not suggest that the weakening of the productivity trend is attributable in part to changes in the socio-political environment that are of recent origin or that have cumulated to a 'critical mass'" (p. 4).

The suggested mode of operation of these factors is, first, through a decline in the rate of capital deepening; second, through a decline of worker effort symptomized by absenteeism and by a drop in hours worked relative to hours paid; third, by a disinclination for risky, innovatory effort, whose manifestation is the observed slowdown in the residual measures of total factor productivity growth; and fourth, through the diversion of resources to regulatory compliance, the benefits of which do not register in measured output even when they should.

These sources of retardation whether great or small—the "suspects," as Fellner calls them—are arguably associated with characteristic features of our mixed economy, even if they are not exclusively due to them. The first of those features is the widening difference between before- and after-tax marginal rates of return to work, saving, investment, and risk taking. The magnitude of the rise in these rates is indicated by the overall increase of total government expenditures from 20 percent of *GNP* in 1947–49 to 28 percent in 1963–65 and again to over 32 percent in 1977–79.[13] The incentive effects of the tax increases are still imperfectly understood, but there is little reason to suppose they are not distinctly unfavorable.[14] Allied to the effects of rising tax burdens is the possible effect of the cumulating "social security wealth" of individuals on savings and that of other insurance and income-support programs on work.[15] Next, there are the effects of burgeoning regulatory activity. These go beyond the direct resource costs of compliance already mentioned. There are also indirect costs and risks of obtaining administrative and judicial clearance for new projects, the diversion of *R&D* expenditure to meet environmental and safety standards, and the hazards of possible future changes in regulatory requirements. Finally, there are the manifold effects of erratic and accelerating inflation.

[12] Some representative references which illustrate the variety and uncertainty of the results obtained by different investigators are: Edward Denison, especially ch. 9; J. R. Norsworthy, Michael Harper, and Kent Kunze, pp. 387–421, and the accompanying discussion and reports by Peter Clark, Martin Baily, Denison, and Michael Wachter; Laurits Christainsen and Haveman; Robert Coen and Bert Hickman; Kendrick (1980); M. Ishaq Nadiri.

[13] See *Economic Report of the President* (1980, Table B-72).

[14] See James Tobin, Lecture III.

[15] The large effect shown in Feldstein's original, much-noticed time-series analysis (1974) has been thrown into doubt by the discovery of a flaw in his computer program. In a forthcoming NBER working paper, he now finds a smaller but still significant effect. Such time-series estimates remain uncertain because it is hard to measure expected Social Security benefits and hard to separate the effects of Social Security wealth on saving from those of other variables during periods of relative stability, as in samples covering the postwar years alone. The conclusion that Social Security benefits work to reduce saving, however, is supported by other studies, based on samples of individual households and on cross-country evidence, to which Feldstein refers in his new working paper.

Inflation belongs in this litany because our pursuit of alternative welfare goals has thus far also involved a tolerance, indeed a pressure, for chronic budgetary deficits, and an understandable political incapacity to employ monetary and fiscal restraint forcefully and consistently at the risk of elevated unemployment. Inflation, in conjunction with tax rules and accounting practices designed for a stable price regime, has meant very high marginal taxes on returns to capital. In the judgement of some public finance experts, it has also meant a differential burden on business investment compared with that on household borrowing, spending, and investing.[16] If there are fears of accelerated inflation in the future, they carry the prospect of still higher taxes and lower returns while the erratic nature of rapid inflation makes the future more difficult to discern and increases the sense of risk. And if the same fears give rise to a vision of price controls, the risks of investment and innovation are compounded. In any event, inflation compels—or threatens to compel—governments to reduce capacity utilization below its potential. Therefore inflation acts to diminish one of the inducements to invest, as the 1980 business contraction following on financial disorder illustrates. We should remember, moreover, that there is an element of vicious circularity in this aspect of our present conjecture. Inflation has deleterious effects on productivity growth—and unexpected declines in productivity growth exacerbate inflation.

This range of considerations leads some students to the view that the pursuit of alternative welfare goals accounts for a very considerable part of the retardation. Fellner, whom I mentioned before, suggests that "the causes of at least 1 percentage point annual slackening of the trend in output per worker's hour can be found among the 'suspects'" (p. 10). That loss is equal to one-half the observed difference between the private-sector productivity growth since 1973 and that during the quarter century between 1948 and 1973.

Such numbers and the argument that leads to them should be understood to be no more than what they are—a *prima facie* indication that something very substantial may be involved in the choices we make between productivity growth and alternative welfare goals. I would not mention them if I did not fear that there is much to the problem, if not as a cause of the recent abrupt retardation, then as a longer-term secular constraint. Yet, at the present time the argument is only speculative, and the estimated loss still more so. The theoretical and quantitative issues are unsettled and deserve our most urgent attention.[17]

[16] See Feldstein and Lawrence Summers. This study measures the extra taxes imposed by inflation on corporate income both at the level of the corporations themselves and at those of the households and institutions which receive dividends and interest payments or have an equity interest in the corporations. They find that the combined excess tax due to inflation averaged only 16.4 percent of corporate income tax from 1954 through 1968, but rose to an average of 52 percent from 1969 through 1977. As a result, the reduction in the effective combined tax on corporate income, which had been accomplished by the tax acts of the early 1960's, was reversed. The combined tax, which had fallen as low as 55 percent of real corporate income from 1962 through 1967, rose to an average of 68 percent from 1968 through 1977. This somewhat exceeded the rate of the latter 1950's, when the combined tax averaged 65 percent of corporate income from 1954 through 1961.

There is a presumption, though no direct proof, that the increase in the effective tax rate reduces the real after-tax rate of return on capital and, therefore, the rate of business capital formation. Feldstein and Summers also argue that, since the impact of inflation on taxes works unevenly, it makes for capital misallocation among industries, encourages more investment in inventories and less in equipment and structures, and tends to shift investment away from the corporate sector and towards residential construction and consumer durables (pp. 47–48). See also Patric Hendershott. Inflation, in conjunction with the tax system also works to increase real tax rates on forms of income other than capital, but this effect is relatively small. See Stanley Fischer and Franco Modigliani (pp. 10–11) which provides a general discussion of the costs of inflation.

[17] In particular, it is possible to propose calculations of the effect of cyclical or other transient changes in the intensity of resource use which suggest that no underlying slowdown occurred before 1973. (See Denison, chs. 7–9.) On such a view, there is a strong suggestion that our troubles do not lie in any generalized impact of the welfare and regulatory programs of government, but are mainly confined to the effects of two developments which either occurred or intensified

VI

So much for the past. We must now try to look ahead. What general view of the future is it sensible to entertain? And what are its implications?

Since our understanding of the productivity retardation of the last dozen years is so clouded, conjecture about the future must be still more fuzzy. True, the negative impact of the recent big influx of inexperienced young workers is due to be reversed. In looking ahead, however, more basic questions need to be addressed. No one, indeed, ought to doubt the persistence of some substantial continuity in what Solomon Fabricant has identified as

> the basic factors underlying economic growth in the United States: the tastes and preferences of the American people, the economic opportunities and alternatives open to them, the social framework within which they live and work together, and the relations of the United States with the rest of the world. Different assumptions would be contrary to all experience and could only lead to wild speculation. [So he concludes] The trend of national output per worker-hour will...continue to be upward. [p. 1]

I agree; but, as Fabricant also asks, how fast will the trend line rise? A "substantial degree of continuity" is not the same as ironclad fixity, and much of this talk has already pointed to some change in Fabricant's basic factors. Within the country, preferences and goals have changed in the degree to which concern for income security, equality of opportunity, environmental protection, and consumer and worker safety sways votes and, to some degree, personal behavior. Corresponding to these shifts in tastes and concerns, the "social framework within which we live and work together" has been recast. The government has come to play a larger role in shaping the "economic opportunities and alternatives" open to us—while imposing burdens on our growth potential whose weight we can now suspect but cannot yet clearly assess. Partly because of higher incomes, partly because of changes in industrial and labor market organization, and partly because of government regulation and income support, there has been a decline in market flexibility—in the responsiveness of prices and wages to the balance of supply and demand, and of people's own responsiveness to price changes—the implications of which Tibor Scitovsky sketched last year.

Our relations with the rest of the world have also changed in ways which I believe are dominantly, but not entirely, unfavorable to *U.S.* growth prospects. The economic rise of Europe and Japan has, indeed, brought those countries to the technological

after 1973, namely, the great increase in the price of energy and the rapid, accelerating, and erratic inflation. Our mixed economy is then implicated to the extent that it works to sustain, if not generate, inflation, and to the extent that our welfare concerns impede the formulation and execution of an energy policy consistent with the maintenance and rapid rise of measured productivity. Continuing work may well clear up these questions about the responsibility of public policy for the current retardation, but, for the time being, we have to live with uncertainty.

The puzzle is still further confused by the experience of the continental European countries. Their fiscal burdens are on the whole heavier than those of the United States, yet their productivity retardation does not generally begin before the oil shock of 1973–74 and the aggravated inflationary disorders that followed. One must, therefore, ask whether the longer persistence of high European productivity growth rates did not reflect a difference in "cyclical" experience. Unlike the United States, they did not generally enjoy a cyclically induced intensification of resource use in the early 1960's and, therefore, a cyclical acceleration of productivity growth. They had no occasion, therefore, to suffer a cyclical retardation in the latter 1960's, as the United States may have done as our economy approached capacity utilization. We may also ask whether the Europeans were more resistant to the incentive effects of heavy taxes and large transfers because of the special factors supporting their great postwar growth booms; or perhaps because their tax and transfer systems are designed differently than ours; or perhaps because of still other matters that differentiate their economies and societies from our own. Or is it the case that what I have referred to as the "suspect factors"—other than inflation and monetary disorder itself—have little to do with the observed retardations of productivity growth? Clearly, the theoretical and empirical issues embodied in these questions call for our very urgent attention.

frontier in many fields. On that account, the effort and experience on which world technological advance rests now has a wider base. The United States, therefore, should now begin to profit more from other countries' technical effort even as other countries borrow from us. It remains to be seen, of course, whether we shall prove as successful at borrowing and adapting foreign technology as some other countries have been.

The advance of other countries, however, also has a darker side for us. The development of many industries in which this country has long been a leader is now threatened by the competition of other countries. This changes the prospects for *U.S.* productivity growth to our disadvantage. It is harder for an industry to push forward, or even to keep up with, the technological frontier when its rate of expansion slows down, still harder when it is contracting. It is an old story that, in the course of aggregate productivity growth, the rise of new, more rapidly progressing industries constricts the growth of the old. That is Schumpeter's "creative destruction," and it helps explain why retardation in the growth of output and productivity is the normal fate of individual industries within a country, while the growth rate of the aggregate remains constant or even speeds up. The reverse, however, is not necessarily true, nor even probably true. We cannot count on new, more progressive sectors stepping into the breach merely because the development of our old industries is constricted by foreign competition. Foreign success, of course, offers us cheap imports. Yet the experience of Britain from 1870 to 1913 presents this country with a worrisome historical question mark. As Britain's basic industries lost their leadership and markets to the United States, Germany, and other countries after 1870, Britain's labor productivity growth rate was halved compared with previous decades, and her average total factor productivity growth during the forty years after 1870 fell to zero.[18] The question is: Can we mount a more energetic and successful response to the challenge of newly rising foreign competitors after 1970 than Britain did after 1870?[19]

The relative decline of *U.S.* economic and political power carries with it other disadvantages, and not for ourselves alone. The leadership of the United States in the liberalization and stabilization of international economic relations was one of the bases for rapid world-wide productivity growth in the postwar years. We were able to assert that leadership because superabundant economic strength permitted us to propose arrangements beneficial to ourselves but generous to other countries, and because dominant political power persuaded sometimes recalcitrant partners to cooperate. Today, with *U.S.* influence reduced and *U.S.* as well as European industries under pressure, the world economy is threatened by a resurgence of protectionism, in which this country is itself taking part. The world-wide price discipline, which a relatively stable *U.S* monetary policy imposed through the dollar-exchange standard, has, for the time being, been lost. And with *U.S.* influence diminished, effective international cooperation in the petroleum market and in other aspects of relations between industrialized and developing countries has been beyond our reach.

In these circumstances, it is just as difficult to maintain a vision of an unbroken 3

[18] The following figures support these statements. All are compound growth rates per year.

	1856–73	1873–1913
(1) Gross domestic product	2.2	1.8
(2) Man-hours	0.0	0.9
(3) Labor input adjusted for quality	1.4	1.7
(4) Output per man-hour	2.2	0.9
(5) Output per unit of quality-adjusted labor input	0.8	0.1
(6) Total factor productivity	0.6	0.0

Source: R. C. O. Matthews, C. Feinstein and J. Odling-Smee. Line (1), Table 16.1; lines (2) and (3), Table 16.4 (quality adjusted for age, sex, length of schooling, and intensity of work associated with number of hours); line (4) = line (1) – line (2); line (5) = line (1)–(3); line (6), Table 16.2 based on total factor input with labor input adjusted for quality.

[19] The British experience, of course, presents a prior question. Which came first, the successful competition of the younger industrial countries in Britain's basic industries, or her own loss of dynamism? Britain in those years was suffering from more than foreign competition in world markets, but my argument makes that competition partly responsible for Britain's national economic retardation (compare Matthews, Feinstein, and Odling-Smee, ch. 17).

percent trend rate of private-sector productivity growth as it is to discard a vision of a trend rate which continues to be significantly positive. It should, therefore, be no surprise that official and other responsible projections foresee productivity growth rates that lie above zero, but significantly below the average postwar rate.[20]

The uncertainty surrounding any such forecasts can hardly be overstated. The progress of science and the enlargement of the knowledge bases of technology go on apace. Our problem is to overcome or mitigate the forces that are checking our ability to give our growing knowledge practical application and to exploit its benefits fully. There are both physical and monetary sides to our present condition which make our prospects particularly perplexing. On the physical side is the new energy question. Quite apart from the policies we pursue—which may themselves be of crucial importance—we do not now know on what terms supplies will be available, even so far as they depend only on physical and technological considerations. We are uncertain about the elasticity of substitution between energy and other resources, and we do not know how much technological progress will itself be impeded as we try to move along a less-energy-intensive path than we have followed in the past. The spread of industrialization from Europe and North America to Asia and Latin America also raises questions about the supplies of other primary materials. As for money, so long as we prove incapable of overcoming our present disposition to inflation, we shall not be able to reach and exploit what would otherwise be the growth potentials of our economy. But if we ever do regain a substantial degree of price stability, we may be happily surprised, even as the Stagnationists of the 1930's were astonished by our growth performance in the postwar period.

VII

In spite of these uncertainties and whatever pleasant or gloomy surprises they may hold, we can hardly avoid the present presumption that our policy choices in the calculable future will need to be made in a less favorable growth environment than that of the generation just past. Our problem of choice will be all the more aggravated if, as now seems likely, the burden of defense expenditures must increase.

That means, first, that our further pursuit of social welfare goals will have to be paid for out of smaller increments of output and income. So, there will be a more difficult problem of choice even if our growth rate itself were not affected by what we choose. It means, second, that the impact of our choices on the measured growth rate itself becomes a more pressing concern and may go far to determine whether the projections now entertained are, indeed, ratified by history or belied. The new, more confined growth environment means, third, that the role of government as a contributor to measured productivity will also be more vitally important, not merely insofar as the government may act to minimize its regulatory or fiscal impact on private performance, but also in the support it gives to research, education, information, labor mobility, and to human capital formation generally.

As we think about these questions, we should not be trapped in the grooves of popular debate. As already said, the alternative paths to economic progress do not present us with clear-cut choices between welfare through government production guidance and income redistribution on the one side, and welfare through private productivity growth on the other. Even if we cared for little except the private use of private earnings, we could not ignore the costs and conflicts arising from the economic and social displacements which accompany growth. We could not, for

[20] For example, in its 1979 *Economic Report of the President*, the Council of Economic Advisors estimated the current trend rate of advance of labor productivity in the national economy at 1.5 percent a year, corresponding to 1.75 percent in the private sector, which is little more than half the postwar pace. In its 1980 *Report*, moreover, the Council writes: "Since the average rate of increase during the past 6 years has been below that figure [of 1.5 percent], the trend rate of increase [in the national economy] may very well be still lower, perhaps 1 percent" (p. 88). For further discussion and other projections, see Fabricant (pp. 63 ff.).

example, disregard problems which the changing structure and role of the family bring in their train. The state of our cities with all their problems of poverty, crime and deteriorating education, and all their exposure to the pressures of racial concentration and frustration, should be a sufficient reminder. All are bound up with the productivity growth process itself. They are sources of antagonism, conflict, and decline of personal quality which will work to constrain growth unless moderated.[21]

VIII

In the new, less-favorable growth environment, the tensions between productivity and other welfare goals are screwed several notches tighter. The success of our mixed economy and pluralistic society in the next generation will depend heavily on how those tensions are managed. In present circumstance, therefore, economic progress turns very largely on the policies we pursue, on what we do through government, and how we do it. As things now stand, however, we can hardly be said to be adopting policies so much as floundering among them, recoiling from growth and backlashing against the recoil, for lack of knowledge and for lack of proper political institutions to use such knowledge as we have.

The gaps in our knowledge define the job for economics. Virtually every facet of the way productivity depends on policy involves matters of fact still to be established. What is the elasticity of substitution between energy and other resources, and how much will it cost us in future output if we forego the cheapest mode of increasing energy supplies in order to provide a greater degree of protection for environment and people? What are the full benefits and what are the full costs of other environmental or safety measures as now legislated and applied? And how much could we save if we sought similar levels of protection more efficiently by making larger use of market incentives as regulatory devices? What are the effects of different levels and—just as important—different types of taxes and transfers on the supplies of saving, investment, and risky enterprise, and on the supply of labor and the quality of people? What is the full range of our government expenditure which has the character of capital formation—and what are the returns to investment in education and in research and development? What would our progress in productivity look like if we tracked it by a system of national accounts more relevant to long-term change in economic welfare than our conventional national product? The questions go on and on. These are matters to which, for the most part, economists have only recently turned. They are now being attacked with vigor, which is testimony to the fact that the aggravated tension between measured productivity growth and other welfare goals is eliciting a constructive response. There are promising beginnings of useful analytical and empirical work, and these will benefit from future experience and experiment. At the same time, our knowledge about this entire range of questions continues to be uncertain.

The weakness of our knowledge, moreover, is matched, probably exceeded, by the weakness of the political institutions and procedures through which that knowledge must be brought to bear. The structure of government and politics, which served us well enough during a more individualistic era and before the population movements of the last fifty years, has not been successfully adapted to the new scale and complexity of public functions. Let me just allude to three political problems.

One concerns federal budget procedure. In principle, the budget is the place where the conflicting claims of special interests should confront, not only one another, but also the general interest in economy and in maintaining a balance between private and public uses of income. It is also the place where our concern for increasing welfare by

[21] This, however, does not mean that our present welfare and training programs are uniformly effective emollients and remedies for the dislocations and maladjustments of growth. Nor does it mean that our present income-support programs may not, in some instances, have little-understood, deleterious side effects on family life and individual quality. Nor does it mean that we now know how to do better.

raising measured productivity should be brought into balance with our interest in other welfare goals. But our budgetary process, in spite of improvements in recent years, remains weak. Tolerance for deficits is the overt, inflation is the covert, mode by which competing claims are reconciled. For lack of a systematic way of facing the future costs of present acts, three-quarters of the budget consists of "uncontrollable" items. Capital investment is not distinguished from current consumption. We have just begun to recognize that regulatory acts impose private costs of compliance, analogous to excise taxes, which must somehow be brought within the budgetary ambit of the public household.

A second matter is what, by pleasant euphemism, is called our system of local government. Fractionated geographically and functionally and poorly coordinated, operating in a confused relation to the federal government, plagued by financial crisis reflecting in part the disjunction between the populations they serve and the tax bases on which they rest, our towns, cities, and districts are fertile generators of external costs, duplicative and costly regulation, and chronic neglect. If, as historians generally agree, Britain could not have carried through its Industrial Revolution without the great Victorian reforms of local government, we ought to be asking whether we can meet the emerging problems of growth and welfare in the second half century of our mixed economy without also facing up to the need for systematic local government reform.

The third matter is both basic and diffuse, and that is the weakness of our party system. It is a commonplace that our national parties are no more than fluid, transistory, and undisciplined coalitions of regional and economic interest groupings. Their lack of central organization and authority, reflecting the size and diversity of the country and people, and our lack of ideological commitment, lays us wide open to the distorting influence of special-interest lobbies and single issue politics. In our political life, we are all too vulnerable to particularistic pressures and all too resistant to the needs of general interest legislation.

IX

The rationale supporting the development of our mixed economy sees it as a pragmatic compromise between the competing virtues and defects of decentralized market capitalism and encompassing socialism. Its goal is to obtain a measure of distributive justice, security, and social guidance of economic life without losing too much of the allocative efficiency and dynamism of private enterprise and market organization. And it is a pragmatic compromise in another sense. It seeks to retain for most people that measure of personal protection *from* the state which private property and a private job market confer, while obtaining for the disadvantaged minority of people *through* the state that measure of support without which their lack of property or personal endowment would amount to a denial of individual freedom and capacity to function as full members of the community.

The viability, to say nothing of the success, of this compromise demands a rough, three-cornered balance between the degree to which we look for economic progress through the development of our powers of production by private action, the degree to which we try through government to protect and promote those aspects of production which markets do not reach, and the degree to which we use governments to alter and cushion the market's income verdicts and to resolve the social conflicts which are inherent in growth and change. Until recently, we have paid inadequate attention to the requirements of achieving that balance wisely. We were able to neglect the problem because we enjoyed the amplitude of a run of fortunate years, when rapid and steady growth was the unseen moderator of the tensions of balance. In the new and less favorable environment of growth, however, the tensions between productivity and the alternative dimensions of welfare are aggravated and the problems of balance—of how much to do and how to do it—are more severe.

In the last analysis, values—feelings, tastes, and sympathies—control choices. But those feelings and sympathies should not have to be deployed with the sad deficiencies of knowledge which, in so many spheres, is the case today. Nor should we have to bring feelings and knowledge to bear through political institutions and procedures which are as imperfect as those through which we now act.

When Keynes spoke of the potential efficiency of a "wisely managed" capitalism, he was assuming that the knowledge necessary for wise management was either in hand or would be forthcoming. But he did not seem to be thinking about the limitations of the political process in bringing knowledge to bear. Now that economists and other social scientists have begun to work at it, we can be cautiously hopeful that our knowledge about both the tradeoffs and the complementarities between productivity growth and the other dimensions of economic welfare will gradually improve. For the calculable future, however, our limited political capabilities may well prove to be the most binding constraint on our ability to work out a social organization which, as Keynes said, "shall be as efficient as possible without offending our notions of a satisfactory way of life."

Contemplating these obdurate realities, what can one say to conclude this talk on an upbeat note? The best I can do is a somewhat inspirational passage from a lecture by Jacob Viner, who, as we all know, was no flaming New Dealer, no Great Society man, and no Keynesian. I am fond of this passage, not only because of its sturdy determination, but also because it displays so well Viner's precise but involuted mind, and his amiable weakness for the nonstop sentence. At the close of a long critique of the American welfare state, which is the mixed economy I have been talking about,

Appendix Table 1—Growth Rates of Productivity (Output per Hour) in the Private Sector: Measures Across Phases of Depression or Stagnation, and Phases of Prosperous Development, 1892–1979

	Growth Rates of Productivity[a]		Deviations of Cross-Stagnation Rates from Neighboring Phases of Development		
	Across Depression or Stagnation Phases[b]	Across Phases of Prosperous Development[b]		Absolute Differences[c]	Percentage Differences
1892–99	1.47		1892/99		
			–1899/1907	–0.55	–27.2
1899–1907		2.02			
			1907/13		
1907–13	1.26		–1899/1907	–0.76	–37.6
1920–29		2.76	1929/37		
			–1920/29	–1.11	–40.2
1929–37	1.65				
			1929/41		
1929–41	2.51		–1920/29	–0.25	–9.1
1948–65		3.2	1965/79		
			–1948/65	–1.6	–50.0
1948–73		2.9			
1965–79	1.6		1973/79		
			–1948/65	–2.6	–81.2
1973–79	0.6		1973/79		
			–1948/73	–2.3	–79.3

Sources: 1899–1941, Kendrick (1961, Table A-XXII); 1948–79, see fn. 8.

[a] Shown in percent per year.

[b] Terminal years of phases are NBER business cycle peaks, except 1965.

[c] Shown in percentage points.

Viner says:

> For all these reasons,...there is in the abstract no reason for making an idol of the welfare state in its American form or for dedicating ourselves unreservedly to its continuance as it is today without qualification or amendment. Given the...imperfection of the procedures whereby it deals with problems which it cannot evade or defer or with problems which special interests may press upon it for premature resolution, it would be only by the dispensation of a benevolent Providence that it would ever make precisely the right decisions or always avoid major mistakes. It does not have theoretical superiority over all conceivable alternative systems.... If...I nevertheless conclude that I believe that the welfare state, like old Siwash, is really worth fighting for and even dying for as compared to any rival system, it is because, despite its imperfections in theory and in practice, in the aggregate it provides more promise of preserving and enlarging human freedoms, temporal prosperity, the extinction of mass misery, and the dignity of man and his moral improvement than any other social system which has previously prevailed, which prevails elsewhere today or which, outside Utopia, the mind of man has been able to provide a blueprint for.
>
> [pp. 166–67]

APPENDIX TABLE 2—INDICATORS OF CHANGE IN MATERIAL WELFARE

	Compound Growth Rates (percent per year) 1948–65	 1965–73	 1973–79
Productivity and Per Capita GNP			
(1) *GNP* per employed worker	2.57	1.60	0.25
(2) Workers per capita	−0.42	1.02	1.43
(3) *GNP* per capita	2.14	2.64	1.69
Real Disposable Income per Capita			
(4) All income (*PCE* deflator)	1.90	3.22	1.75
(5) ——— (*CPI* deflator)	2.21	2.85	0.84
(6) All income less transfers (*PCE*)	1.74	2.55	1.29
(7) ——— (*CPI*)	2.05	2.18	0.38
(8) All income less transfers and other labor income[e] (*PCE*)	1.58	2.27	0.79
(9) ——— (*CPI*)	1.89	1.90	−0.11
Workers' Earnings			
(10) Real compensation per full-time equivalent employee (*PCE*)	2.66	2.69	0.84[a]
(11) ——— (*CPI*)	2.96	2.32	0.19[a]
(12) Real wages and salaries per full-time equivalent employee (*PCE*)	2.35	2.20	0.10[a]
(13) ——— (*CPI*)	2.66	1.83	−0.54[a]
(14) Real wage and salary income, full-time white males (*PCE*)		3.01	−0.41[b]
(15) ——— (*CPI*)		2.61	−1.03[b]
Median Real Total Income, Persons 14-Years Old and Over			
(16) All males (*PCE*)	2.44[c]	2.00	−1.09[a]
(17) ——— (*CPI*)	2.65[c]	1.64	−1.73[a]
(18) Year-round full-time male workers (*PCE*)	2.61[d]	3.03	−0.32[a]
(19) ——— (*CPI*)	2.81[d]	2.66	−0.95[a]
Median Real Total Family Income			
(20) *PCE* deflator	2.74	2.99	0.48[a]
(21) *CPI* deflator	3.05	2.62	−0.16[a]

(Continued)

APPENDIX TABLE 2— (Continued)

	Percentage of Persons Living in "Poverty"			
	1965	1968	1972	1976
Pretransfer				
(22) Official measure	21.3	18.2	19.2	21.0
(23) Adjusted official measure	–	18.0	18.2	21.1
(24) Relative measure	21.3	19.7	22.2	24.1
Posttransfer				
(25) Official measure	15.6	12.8	11.9	11.8
(26) Adjusted official measure	–	10.1	6.2	6.5
(27) Relative measure	15.6	14.5	15.7	15.4

Sources: Lines (1) *Economic Report of the President*, (*Report*), January 1980, Table B-2, col. (1) (1979, rev.) and Table B-27, col. (2)+col. (4); (2) *Report*, Tables B-27, col. (2)+col. (4) and Table B-26; (3) *Report*, Tables B-2 and B-26; (4) *Report*, Table B-22, col. (4); (5) *Report*, Table B-22, col. (3) deflated by *CPI*, Table B-49, col. (1); (6) *Report*, Table B-22, col. (1) less Table B-20, col. (14), divided by population, Table B-26 and *PCE* deflator, Table B-3, col. (2); (7) See line (6), except *CPI* deflator, Table 49, col. (1); (8) Disposable personal income less transfers current $, as in line (6) less "other labor income," *Report*, Table 20, col. (8) and deflated for population and prices as in line (6); (9) See line (8), except *CPI* deflator as in line (7); (10) and (11) *Survey of Current Business*: *A Supplement*, *The National Income and Product Accounts of the United States*, *1929–1965*, Tables 6.1 and 6.4 and corresponding Tables for *SCB*, July 1977 and 1979, deflated by *PCE* and *CPI*, respectively; (12) and (13) *Survey of Current Business*, *1929–65*, Table 6.5 and corresponding tables in *SCB*, July 1977 and 1979; (14) U.S. Bureau of the Census, *Current Population Reports*, Series P-60, with *PCE* deflator, as in line (6), above; (15) Same, with *CPI* deflator, as in line (5), above; (16)–(19) Same, Series P-60, No. 120, Table 14, with *PCE* or *CPI* deflators, as indicated, (20) and (21) *Current Population Reports*, Series P-60, No. 120, Table 3, deflated as in lines (16) to (19), except 1948 from *Report* January 1980, Table B-25, converted to 1972 dollars; and (22)–(27) Robert Plotnick and Timothy Smeeding, "Poverty and Income Transfers: Past Trends and Future Prospects," *Public Policy*, 27, No. 3 (Summer, 1979), Table 1. The official measures count the number of persons living below constant real (that is, inflation-adjusted) poverty lines defined for households with different characteristics. The adjusted figures correct the official figures for underreporting of income and, in the posttransfer estimates, for direct taxes and for receipts of transfers in kind. To obtain the relative measures, the authors "set the relative poverty lines equal to the federal ones [in 1965]. In succeeding years, the relative lines are increased at the same rate as the median income" (p. 258).

Notes: *PCE*=implicit *GNP* deflator for personal consumption expenditure; *CPI*=Bureau of Labor Statistics Consumer Price Index, all items.

[a] 1973–78

[b] 1973–77

[c] Average 1947 and 1950 to 1965

[d] 1955–65

[e] "Other labor income" includes "employers contributions to private pension, health, unemployment, and welfare funds; compensation for injuries; director's fees, pay of the military reserve; and a few other minor items."

REFERENCES

A. F. Burns, "The Condition of the American Economy," *The Francis Boyer Lectures on Public Policy*, Washington 1979.

G. B. Christainsen and R. H. Haveman, "The Determinants of the Decline in Measured Productivity: An Evaluation," paper presented at a joint session of the Society of Government Economists and the American Economic Association, Atlanta, Dec. 1979.

R. M. Coen and B. G. Hickman, "Investment and Growth in an Econometric Model of the United States," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 214–19.

S. Danziger, R. Haveman, and R. Plotnick, "Income Transfer Programs in the United States: An Analysis of their Structure and Impacts," prepared for the Joint Economic Comm., mimeo., 96th Cong., 1st sess. 1979.

Edward F. Denison, *Accounting for Slower Economic Growth*, Washington 1979.

R. Eisner, "Total Income, Total Investment, and Growth," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 225–31.

Solomon Fabricant, *The Economic Growth of the United States*, Montreal; Washington 1979.

W. Fellner, "The Declining Growth of American Productivity: An Introductory Note," in his *Contemporary Economic Problems*, Washington 1979, 3–12.

M. Feldstein, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 905–26.

______, "Social Security, Induced Retirement, and Aggregate Capital Accumulation: A Correlation and Updating," Nat. Bur. Econ. Res. work. paper, no. 583, Cambridge, Mass., forthcoming.

______ **and L. Summers,** "Inflation and the Taxation of Capital Income in the Corporate Sector," Nat. Bur. Econ. Res. work. paper no. 312, Cambridge, Mass., Jan. 1979.

S. Fischer and F. Modigliani, "Towards an Understanding of the Real Effects and Costs of Inflation," Nat. Bur. Econ. Res. work. paper no. 303, Cambridge, Mass., Nov. 1978.

P. H. Hendershott, "The Decline in Aggregate Share Values: Inflation and Taxation of the Returns from Equities and Owner-occupied Housing," Nat. Bur. Econ. Res. work. paper no. 370, Cambridge, Mass., July 1979.

John W. Kendrick, *Productivity Trends in the United States*, Princeton 1961.

______, "Discussion [on Denison]," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 232–33.

John Maynard Keynes, *The End of Laissez-Faire*, London 1926.

S. S. Kuznets, "Driving Forces in Economic Growth: What Can We Learn from History," *Proceedings of the 1980 Kiel Conference on Economic Growth*, forthcoming.

R. C. O. Matthews, C. Feinstein, and J. Odling-Smee, *British Economic Growth*, Stanford 1981.

M. I. Nadiri, "Sectoral Productivity Slowdown," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 349–52.

J. R. Norsworthy, M. J. Harper, and K. Kunze, "The Slowdown in Productivity Growth: Analysis of Some Contributing Factors," *Brooking Papers*, Washington 1979, *2*, 387–421.

R. Plotnick, "Social Welfare Expenditures: How Much Help for the Poor?," *Policy Analysis*, Summer 1979, *5*, 271–89.

T. Scitovsky, "Can Capitalism Survive?—An Old Question in a New Setting," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 1–9.

James Tobin, *Asset Accumulation and Economic Activity*, Chicago; Oxford 1980.

J. Viner, "The United States as a 'Welfare State,'" in Edgar O. Edwards, ed., *The Nation's Economic Objectives*, Chicago; London 1964, 151–67.

Burton A. Weisbrod, *The Economics of Poverty, An American Paradox*, Englewood Cliffs 1965.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1964; 1979; 1980.

U.S. Social Security Administration, Office of Research and Statistics, "Research and Statistics Note No. 15," Washington, Dec. 29, 1978.

Competition and Unanimity

By HARRY DEANGELO*

If the investment and financing decisions of a given firm are perceived as affecting owners' consumption opportunities only through their impact on personal wealth, then firm value maximization is unanimously supported because it maximizes every owner's wealth and, therefore, consumption opportunities and utility. This opportunity set dominance argument is the essence of the Fisher Separation Theorem (*FST*).[1] Eugene Fama and Merton Miller (pp. 176–78) and others have emphasized that the basic logic of the argument applies in economic environments considerably more complex than the certainty world in which it was originally formulated. Their cogent observation notwithstanding, there is an extensive and growing literature concerned with specifying mathematically complex conditions for unanimity in economies with incomplete risk-sharing capabilities. Ironically, many of the papers comprising this literature appear to disagree regarding the critical market conditions for unanimity. For example, even the synthesis papers in this area differ in their characterizations of the competitive conditions in incomplete markets which they claim to be necessary and sufficient for unanimity (for example, see David Baron, 1976a, 1979; Sanford Grossman and Joseph Stiglitz; Hayne Leland, 1973, 1977; Frank Milne, 1974; Niels Nielsen).[2]

*Assistant professor of finance, University of Pennsylvania. This paper is a substantially revised version of the first chapter of my dissertation. I gratefully acknowledge the help of my reading committee (J. Hirshleifer; C. G. Krouse, Chairman; K. V. Smith), and of H. E. Leland, E. M. Rice, and J. F. Weston. I also benefited from discussions with N. F. Chen, L. Y. Dann, L. E. DeAngelo, R. W. Masulis, D. Mayers, E. Omberg, and S. A. Ross. Participants in the finance workshops at University of California-Los Angeles and the Universities of Chicago, Pennsylvania, and Washington provided useful comments on an earlier version of this paper. Responsibility for remaining errors is mine.

[1] The *FST* states that, in competitive markets, the firm's investment decision can be separated from owners' consumption decisions in the sense that, regardless of the details of preferences, the utility-maximizing investment decision is that which maximizes current firm value. See Irving Fisher (p. 141), Jack Hirshleifer (p. 14), and Fama-Miller (p. 69).

[2] An entire subliterature concerned with *ex post* unanimity analysis (generated by the work of Steinar Ekern and Robert Wilson, and Leland, 1974 has been applied to such diverse areas as international trade theory (see Baron, 1976b) and accounting (see William Beaver and Joel Demski). It claims that non-value-maximizing decisions could be unanimously supported in incomplete market economies. In an important but unheralded paper, Nielsen clearly established the falsity of this claim. According to the *ex post* theorems, unanimously supported corporate decisions exist, provided only that a given firm's decisions have no impact on the economy's risk-sharing capabilities (i.e., regardless of any noncompetitive perceived price impacts). But this "spanning" condition is satisfied in a certainty world in which a given firm has monopoly power over the market interest rate. And unanimity clearly fails under monopoly. Nielsen reconciled this apparent contradiction by explaining that *ex post* analysis does not imply the existence of unanimously supported corporate decisions under monopoly, but rather yields the far weaker result that owners will agree on the direction of change (but generally will disagree as to the magnitude of that change) in a corporate decision variable. In other words, given only spanning (and no price-taking assumption), owners disagree on the optimal level of the decision variable and in this global sense unanimity fails. I shall not impose the assumptions of *ex post* analysis in this paper but will focus instead on *ex ante* analysis. The differences are that 1) *ex post* analysis assumes that all individuals' endowments are utility maximizing while *ex ante* analysis makes no such assumption and 2) *ex ante* analysis invokes some notion of competitive price-taking behavior while *ex post* analysis does not; see Nielsen. Incidentally, Nielsen's general equilibrium concept of the competitive market conditions for *ex ante* unanimity in incomplete markets comes close to those set forth below. However, his general equilibrium concept of competition (labelled actual price independence (*API*), below) is so restrictive that it not only precludes many obviously correct partial-equilibrium unanimity theorems, but also implies unanimous indifference to the decisions of a given firm. Finally, as Leland has pointed out in a private communication, *ex post* analysis may lead to a useful comparative statics theory of firm behavior in economic environments in which (i) unanimity fails in the global sense that owners disagree on the optimal decision level, but (ii) there is agreement in a local neigh-

My basic contention is that the complex analytics employed in the literature have obscured the simple economic logic which underlies unanimity theorems. I present a simple yet general theoretical framework which reconciles the complex web of alternative assumptions invoked in extant unanimity theorems. I demonstrate in Section I that every theorem which establishes unanimity in quite general collective choice settings must rely on simple opportunity set dominance arguments which parallel the standard Fisherian logic. I then apply this result in a market context and formulate a general definition of competition which 1) admits the possibility of incomplete markets, 2) embodies the apparently differing notions of competition invoked in the literature as special cases, and 3) which, by the same logic as the original *FST*, implies that value maximization is the unanimously supported corporate policy (Section II). I then distinguish between unanimity theorems based on partial and general equilibrium notions of market competition. It will be shown that unanimity theorems which establish strict preference (rather than unanimous indifference) over a firm's decisions must rely on a partial-equilibrium notion of competition (Section III). Throughout Sections II and III, I emphasize the essential economic logic of unanimity theorems without recourse to complex utility differentials or other elaborate mathematical arguments often found in the literature. I conclude the paper with a brief summary (Section IV).

I. Collective Choice and Unanimity

This section concisely specifies the essence of the collective choice problem and isolates the general opportunity set dominance condition for unanimity. A collective is a group of two or more individuals who are linked by a common decision potentially affecting the well-being of all. In the choice setting modeled here, decisions may be classified as either personal or collective. Personal decisions are made by a single individual and affect only the welfare of that individual. Collective decisions may simultaneously affect each individual's opportunities for alternative personal decisions and thereby affect the welfare of all individuals. To formalize the collective choice setting, define

y^i = vector of objects desired by individual i

$U^i[y^i]$ = utility function of i which is nondecreasing in personal decision variables y^i

$O^i[\mu^i[z] \mid \pi^i]$ = consumption opportunity set of i

z = vector of collective decision variables

Z = exogenously given set of feasible collective decisions

$\mu^i[z]$ = vector of parameters of i's opportunity set which are (potentially) affected by the collective decision, but not by any individual's personal selection of desired objects

π^i = vector of exogenous parameters of i's opportunity set, that is, parameters which are unaffected by either collective or personal decisions.[3]

A collective decision z^* is unanimously supported if and only if it is utility maximizing for all individuals. Formally, z^* is unanimously supported if and only if z^* is a

borhood around the equilibrium decision (where equilibrium is reached through some social choice mechanism). In this paper, I restrict attention to the narrower question: under what conditions will individuals unanimously agree on the globally optimal decision?

[3] This choice setting is obviously quite general and includes many (but not all) interesting economic environments as special cases. For example, the standard competitive paradigm is easily seen to be a special case by defining Z as firms' possibilities set, z as firms' decision variables, $\mu^i[z]$ as individual i's wealth which is potentially a function of firms' decisions, and π as the vector of prices of desired objects (exogenous by the competitive assumption). In this case, i's consumption opportunity set is simply $\{y^i: y^i\pi \leqslant \mu^i[z], y^i \geqslant 0\}$. Notice that if prices were allowed to vary with z, the same analytical apparatus would capture the standard monopoly model. However, this formulation does not allow collective decisions to enter utility functions directly as, for example, in Leland's (1978) information production-unanimity paper. There is no clear link between opportunity set dominance and unanimity in this case.

solution to (1) for all individuals:

$$\text{(1)} \qquad \underset{z,y^i}{\text{maximize}}\, U^i[y^i]$$

$$\text{subject to} \qquad y^i \in O^i[\mu^i[z] | \pi^i] z \in Z$$

Thus, whether a unanimously supported collective decision exists depends upon the z-solution properties of a set of constrained optimization problems. The solution to any given optimization problem ordinarily depends upon both the objective function and the constraint set. In general then, the occurrence of unanimity depends upon both the specific details of individuals' preferences and the precise way in which each individual's consumption opportunity set varies with alternative collective decisions z.

Now, the interesting question is: under what conditions will unanimity obtain (i.e., will all individuals agree on the optimal collective decision) independent of the specifics of preferences?[4] Theorem 1, below, demonstrates that a necessary and sufficient condition for a particular collective decision, say z^*, to be unanimously supported independent of preferences, is that all individuals' opportunity sets be globally dominant at z^*. Formally, individual i's opportunity set evaluated at z^*, $O^i[\mu^i[z^*] \| \pi^i]$, is *globally dominant* if[5]

(GD1) z^* is feasible (i.e., $z^* \in Z$) and

(GD2) for any feasible consumption bundle y^i, there exists a y^{*i} which is feasible given z^* and which provides at least as much of every desired object as y^i (i.e., for each $y^i \in O^i[\mu^i[z] \| \pi^i]$ with $z \in Z$, there exists a $y^{*i} \in O[\mu^i[z^*] \| \pi^i]$ such that $y^{*i} \geqslant y^i$).

THEOREM 1: *If every individual's opportunity set evaluated at $z^* \in Z$ is globally dominant, then z^* is unanimously supported. If all individuals' opportunity sets are not globally dominant under the same collective decision, then utility functions may be assigned to individuals in such a way that a unanimously supported collective decision does not exist.* (The Appendix provides a formal proof.)

The important implication of Theorem 1 is that all theorems which establish preference-free conditions for unanimity (in collective choice settings as characterized above) must rely on a simple opportunity set dominance argument. It is neither necessary nor desirable to consider complicated utility function differential arguments as has become customary in the unanimity literature. Instead, one need only postulate the specifics of the economic environment and then ascertain whether the fundamental conditions of Theorem 1 are satisfied. Moreover, utility differential proofs are undesirable because their mathematical complexity obscures the intuitive and simple opportunity set dominance argument which must underlie the unanimity result to be established. I demonstrate below the manner in which the complicated unanimity proofs in the literature are reducible to simple opportunity set dominance arguments.

II. Market Competition and Unanimity

I next apply the results developed for the general collective choice setting to establish conditions for unanimity in a market economy under conditions of risk. I specify a general definition of market competition which incorporates the possibility of incomplete markets and which implies a generalization of the Fisherian conclusion that maximization of the firm's perceived net market value is unanimously supported. Footnoted discussions demonstrate that the different

[4] Much of the literature claims to deal with this question, but then focuses analytically on establishing conditions for unanimous agreement on the direction of change in a collective decision variable. For example, many studies have specified conditions under which all owners of a firm agree that more should be invested, but this does not imply that all agree on an optimal *policy* (i.e., on how much more). All of the *ex post* unanimity literature is subject to this methodological criticism (see fn. 2) as is much of the *ex ante* literature.

[5] The importance of opportunity set dominance was noted by Hirshleifer (p. 199) and Gordon Pye in their insightful and overlooked (at least by the unanimity literature) discussions of preference-free conditions for optimal investment decisions. Incidentally, notice that a sufficient but not necessary condition for $O^i[\mu^i[z^*] \| \pi^i]$ to be globally dominant for i is that all consumption bundles feasible under z are also feasible under z^*, i.e., $O^i[\mu^i[z] \| \pi^i] \subseteq O^i[\mu^i[z^*] \| \pi^i]$ for all $z \in Z$.

notions of competition invoked in the literature are special cases of the general definition of competition and, therefore, that existing theorems are special cases of the generalized Fisherian value-maximization theorem.

While all of the substantive results presented here can be generalized to multiperiod worlds in which asset risk is modeled through subjective probability distributions, I shall maintain consistency with the unanimity literature and operate within the standard two date, two good (current and future consumption) state preference model. Certainty prevails at time $t=0$ while the uncertainty at $t=1$ is summarized by a finite number of mutually exclusive and exogenous states of nature such that knowledge of a particular state's occurrence removes all uncertainty. Individuals are endowed with current consumption and ownership shares in firms. Firms produce state-contingent amounts of future consumption using current consumption obtained from individuals as the sole input. In addition to choosing a production plan, firms select a capital structure which consists of a set of securities, that is, a set of tradeable claims against the state-contingent returns from production.

At $t=0$, firms make production and capital structure decisions which result in a well-defined set of securities (patterns or vectors of state-contingent claims) available to individuals. Simultaneously at $t=0$, individuals exchange current consumption and securities, in the process 1) obtaining a bundle of current consumption and state-contingent claims to future consumption as well as 2) providing inputs for firms. At $t=1$, the true state of nature is revealed, firms' technological returns and security payoffs become known, and the payoffs are distributed to the postexchange security holders.

Let us assume that the $t=0$ markets for current consumption and securities operate perfectly in the sense that individual exchange opportunities are characterized by:

ASSUMPTION 1: (Zero information costs) *All individuals costlessly know the state-contingent payoffs and prices of all securities.*

ASSUMPTION 2: (Zero exchange costs) *There are no indivisibilities or other costs involved in trading current consumption and securities. This does not imply the very strong assumption that individuals or intermediaries can costlessly create any type of security. Rather, it implies only that individuals can exchange without cost those securities issued by firms.* (Various forms of intermediation are considered in fn. 6.)

ASSUMPTION 3: (Zero enforcement costs) *There are no costs of ascertaining which state has occurred or of obtaining state-contingent payoffs contracted for through security purchases.*

ASSUMPTION 4: (Short sale feasibility) *Short sales of securities are possible (without margin penalty) but are limited by a constraint against personal default.*

ASSUMPTION 5: (Competitive behavior by individuals) *All individuals believe that and are treated as if their personal portfolio choices have no effect on either the set of available securities or the prices of those securities.*

ASSUMPTION 6: (New financing at $t=0$) *At time* $t=0$, *the original or pre-exchange owners of a given firm share proportionately in net firm value (i.e., each owner receives a prespecified positive fraction of the net of investment outlay value of all securities issued by the firm. This wealth can be allocated to obtain a personally optimal consumption bundle).*

Under Assumptions 1–5, the exchange opportunities of a given individual depend on three parameters: (i) the economy's feasible risk-sharing opportunities which can be represented analytically as the vector space spanned by the state-claim payoff vectors of all securities, (ii) the current market value of each pattern or vector in the feasible vector space, and (iii) personal wealth. The production and capital structure decisions of a given firm can affect an individual's consumption opportunities only through their impact on (i)–(iii).

D 17023

If we eliminate the possibility of short sales, then (i) should be modified by replacing "vector space spanned" with "convex cone defined" and (ii) should be modified by replacing "vector space" with "convex cone." The proof of Theorem 2, below, is unchanged by this modification and Assumption 4 is not required for unanimity. This represents a significant generalization over existing unanimity theorems, all of which have assumed short sale feasibility. However, throughout the subsequent discussion we maintain Assumption 4 and the attendant vector space interpretation to facilitate comparison with the literature.

I shall define a competitive market as one in which exchange opportunities are characterized by Assumptions 1–6 and in which production and capital structure decisions of any given firm

ASSUMPTION 7: *do not affect the state-contingent cash flows of securities issued by other firms.* (No technological externalities)

ASSUMPTION 8: *do not affect the vector space of feasible state-contingent claims and do not determine whether current consumption is available to individuals.* (Spanning)

ASSUMPTION 9: *do not perceptibly affect the per unit market price of current consumption and of each feasible pattern of state claims.* (Perceived price independence)

Assumption 8 asserts that a given firm has no impact on the economy's risk-sharing possibilities. In technical terms, it implies that the market is either always complete or always incomplete in the same way (i.e., the same proper subspace of the total state space is always feasible). In the literature, spanning has been invoked through conditions on the instrument richness of the capital market or by imposing specific restrictions on firms' security supply capabilities.[6] Notice that Assumption 9 requires that Assumption 8 holds. If Assumption 8 is violated so that a firm can determine whether or not a particular state-claim pattern is available, then Assumption 9 is also violated, since the firm can dictate a finite price for that pattern (if made available) or an infinite price (if made unavailable).

Assumption 9 asserts that the unit prices of current consumption and feasible patterns of state-contingent claims are perceived as independent of the decisions of a given firm. Many different conditions which imply Assumption 9 have been invoked in the

[6] For example, market richness conditions which imply spanning (Assumption 8) include (we ignore the possibility of affecting current consumption supplies): 1) (Certainty) there is only one possible state of nature at time 1 (see, for example, Hirshleifer, Fama-Miller, Mark Rubinstein). 2) (Complete market) there are always as many securities with linearly independent payoff vectors as there are states of nature (see, for example, Hirshleifer; Milne, 1974; Joseph Stiglitz, 1969). 3) (Perfect substitutes) there is always a perfect substitute security (or portfolio of securities) for each of the firm's securities issued under all feasible production and capital structure decisions (see, for example, Fama-Miller; Milne, 1975). 4) (Unrestricted costless financial intermediation) individuals can costlessly create on personal account securities with *any* vector of state-contingent payoffs (see, for example, Stiglitz, 1974). 5) (Equivalent costless financial intermediation) individuals can costlessly create securities on personal account with payoffs identical to any security that the firm is *able* to create (see, for example, Fama-Miller; Stiglitz 1969). Less restrictive types of intermediation may also suffice for spanning. For example, Stephen Ross has shown that, under fairly mild conditions, costless access to a call and put option market will allow individuals to complete the market. Firm-specific conditions which imply spanning are far too numerous to catalog here. However, Stiglitz' (1969) risk-free debt condition provides a classic example (also see Kare Hagen and Milne, 1975). Other authors (for example, Nielsen; Baron, 1976a) have simply asserted that Assumption 8 holds while recognizing that many different assumptions can lead to spanning. Note that Assumption 8 is not equivalent to the marginal spanning condition invoked, for example, in Baron (1979), Grossman-Stiglitz, Leland (1973), and Ekern-Wilson. Contrary to the claims of these papers, marginal spanning does not imply that the feasible vector space is independent of a given firm's decisions and therefore is too weak a condition for unanimity. For a simple proof, let X_1 and X_2 denote two linearly independent state-claim payoff vectors of securities issued by firms 1 and 2 which span the feasible vector space Ω. Let ΔX_1 denote the change in X_1 due to a change in firm 1's production decision variable. Marginal spanning asserts that $\Delta X_1 \in \Omega$. Now, if $\Delta X_1 \equiv X_2 - X_1 \in \Omega$, it is clear that firm 1's payoff vector is now $X_1 + \Delta X_1 = X_2$ and the feasible vector space has changed (but marginal spanning is satisfied).

literature.[7] A special case of Assumption 9 is when a firm's decisions have no actual impact on equilibrium prices. The distinction between perceived and actual price independence is crucial to an understanding of the theory and is therefore examined below. Finally, notice that Assumptions 7 and 9 together imply that the decisions of a given firm have no perceived impact on the market values of securities issued by other firms.

We can now establish the fundamental unanimity theorem for competitive markets:

THEOREM 2: *In an economy characterized by Assumptions* 1–9, *production and capital structure decisions which are perceived as maximizing the net value of a given firm are unanimously supported by all pre-exchange owners of the firm.*

The logic of the theorem is straightforward and is identical to that of the original Fisher Separation Theorem. In a competitive market, a firm's decisions can affect individuals' consumption opportunities only through changes in personal wealth. The set of attainable goods (i.e., current consumption and patterns of state claims) and the prices per unit of those goods are fixed—that is, (i) and (ii) but not (iii) are fixed by Assumptions 8 and 9. For any individual, more wealth results in unambiguously greater consumption opportunities so that optimal corporate decisions are those which maximize personal wealth (regardless of the specific form of his utility function). Now, a given firm's decisions are perceived as affecting an individual's wealth only through their impact on his proceeds from the $t=0$ net value of that firm (Assumption 7 implies no cash flow impact on other firms and Assumption 9 implies no perceived valuation impact). And since every original owner shares proportionately in net firm value (by Assumption 6), maximizing net firm value is unanimously supported because it simultaneously maximizes the personal wealth and therefore the consumption opportunities of every owner (by Theorem 1).[8]

The new financing Assumption 6 can, with some qualifications, be relaxed without affecting the theorem. At $t=0$ when the firm is first organized, all owners share proportionately in net firm value and unanimously support net firm value maximization. At $t=1$, several different claimant classes may exist (this depends on the specific unanimously supported ownership structure selected at $t=0$) and claimholders polled at this time may not agree on optimal decisions for the firm. Unanimity can fail because net firm value is no longer shared proportionately by all claimholders and the wealth of different owners will (except by coincidence) be maximized under different decisions. Potential future conflicts among claimholders (even with positive costs of conflict resolution) are perfectly consistent with unanimous support for net value maximization at $t=0$ (i.e., Assumption 3 can be

[7] Suppose that a given vector space is feasible and independent of corporate decisions (the market may be complete or incomplete). Then the market value of each feasible pattern is perceived as independent of the decisions of a given firm (i.e., Assumption 9 holds) if those decisions have no perceived impact on 1) An Arrow-Debreu state-claim price vector which values out tradeable patterns (for example, see Fama-Miller for the certainty case, Hirshleifer for complete markets, Roy Radner for incomplete markets), or 2) Each individual's marginal rates of substitution of current consumption for state claims evaluated at an assumed interior optimum to the individual's portfolio choice problem (for example, see Leland, 1973; Baron, 1976a, 1979 for both complete and incomplete markets), 3) Values of prespecified basis vectors for the feasible space (for example, see Milne, 1974; Grossman-Stiglitz). If the market is complete, then Assumption 9 requires perceived fixity of 1), 2), and 3). If the market is incomplete, Assumption 9 requires perceived fixity of 3), but is consistent with perceived variations in 1) and/or 2) (because there is an infinity of implicit price vectors implying the same market value for each pattern in a given incomplete market). Leland (1973, 1977) and Baron (1976a, 1979) falsely assert that unanimity requires perceived fixity of 2) in incomplete markets. Assumption 9 is the necessary condition, not 2) or 1).

[8] This argument applies to the original owners who have positive pre-exchange ownership shares. Nonowners (those with zero pre-exchange shares) are unanimously indifferent to the firm's decisions because they experience no wealth impact. This point is important because it verifies that nonowners are not willing to pay owners to change from the value-maximizing solution, i.e., all costs and benefits have been internalized in determining that net value maximization is unanimously supported.

relaxed) since that strategy maximizes all t = 0 owners' consumption opportunities. Of course, net firm value at t = 0 will capitalize the potential costs of future conflict resolution associated with the firm's investment and financing decisions.

Moreover, even with multiple claimant classes at t = 1, unanimity need not fail at this time. The technical reason for potential unanimity failure at t = 1 is the existence of externalities among different classes of claimants within the firm. However, if side payments can be negotiated costlessly, then net value maximization is unanimously supported and the theorem is still valid. Applying the logic of Ronald Coase and Fama, any other decisions taken by the firm imply that some individual can be made better off under net value maximization (by increasing his wealth) without harming (reducing the wealth) of others.[9] Alternatively, if old financing is protected by perfect me-first rules or nonexpropriation clauses (see Fama-Miller, p. 151), then no externalities exist among claimants and net value maximization is unanimously supported (even without the possibility of side payments) because it *is* wealth maximizing for all individuals.

Theorem 2 depends critically on the perceived price independence Assumption 9 (and therefore also on spanning Assumption 8 which, as noted earlier, is required for Assumption 9). To establish the necessity of Assumption 9, consider the simple certainty world (one state possible at t = 1). Take current consumption as numeraire and denote the price per unit of future riskless consumption by q. For an owner with wealth W, the maximum attainable units of current (future) consumption occur under those decisions which maximize $W\,(W/q)$. If a given firm has monopoly power so that its decisions are perceived as affecting q, then current consumption opportunities are greatest under the set of decisions which maximizes W, and future consumption opportunities will be greatest under the generally different set which maximizes W/q. Clearly, globally dominant opportunity sets do not exist in this monopoly world and, by Theorem 1, utility functions can be assigned to individuals in such a way that unanimity fails.[10]

III. Partial vs. General Equilibrium Analysis

Theorem 2 is a partial-equilibrium unanimity theorem because it relies on the assertion (Assumption 9) that market prices are perceived as independent of a given firm's decisions. Potential price impacts were simply assumed away by asserting that perceived impacts are absent. No attempt was made to demonstrate that market equilibrium prices were in fact invariant to the decisions of a given firm. A general equilibrium unanimity theorem would obtain if, in addition to Assumptions 1–9, the capital market were characterized by[11] actual price

[9] The same line of argument applied across firms allows relaxation of the no technological externality Assumption 7. With costless negotiation of side-payments, all owners of two or more technologically interrelated firms would agree to maximize the sum of the net firm values.

[10] By the same logic, unanimity fails if the firm has monopoly power in the product markets (i.e., perceived power over the prices of goods and services which are embodied in the dated consumption claims). Other imperfections may also imply unanimity failure. For example, Hirshleifer and Fama-Miller have shown that certain types of transactions costs—differential borrowing-lending rates, borrowing costs, and certain types of capital rationing—will be associated with unanimity failure. On the other hand, unanimity will obtain with other realistic market features (for example, with those personal tax codes which imply that a given firm can affect consumption opportunities at t = 0 only through personal wealth). Finally, recognize that in the extreme case in which there is no trade, unanimity on the globally optimal firm decisions will not generally obtain among heterogeneous owners.

[11] *API* is the notion of competition considered in the production unanimity papers by Nielsen and Rubinstein, and the leverage irrelevancy theorems of Stiglitz (1969, p. 781), Fama-Miller (p. 150), Robert Litzenberger and Howard Sosin, and Baron (1976a). Assumption 9 is exactly analogous to Assumption 5, since under both, the decisions of firms and individuals are perceived as not affecting market prices. *API* imposes a stronger notion of price taking on firms than Assumption 5 places on individuals. While *API* is a special case of Assumption 9, there is nothing inherently superior about unanimity theorems based on Assumption 9 or on *API*. Rather, they are simply different theorems based on weaker or stronger (partial or general equilibrium) concepts of competition.

independence (*API*). Under this condition the decisions of a given firm do not affect the actual unit price of each feasible consumption bundle; that is, in the exchange equilibria associated with different investment-financing decisions by a given firm, the same unit price is assigned to current consumption and to each feasible pattern of state claims.

The distinction between Assumption 9 and *API* is subtle. Perceived price independence is a *partial*-equilibrium *assumption*—that is, Assumption 9 asserts that unit prices are perceived as independent of one firm's decisions. Actual price independence is a *general* equilibrium *result*—prices are fixed because the economy's endogenous valuation process places the same value on each consumption bundle in the exchange equilibria associated with different decisions by a given firm. The set of prices which are asserted to be fixed by Assumption 9 can be consistent with a general equilibrium in the sense that aggregate quantities demanded and supplied at these prices are equated. In this case, Assumption 9 involves a conjectural variation in prices about an equilibrium or, more precisely, Assumption 9 asserts that unit prices are conjectured to remain unchanged if a given firm's decisions were to be changed from their equilibrium levels. On the other hand, *API* involves knowledge (and requires proof) that prices *are* the same in the alternative equilibria associated with different decisions by a given firm.

A general equilibrium version of Theorem 2 can be established by delineating the conditions under which *API* obtains. Unfortunately, the conditions leading to *API* are so restrictive that they imply unanimous indifference to the investment and financing decisions of a given firm. It follows that interesting unanimity theorems (those which establish strict preferences rather than unanimous indifference) must rely on a partial-equilibrium notion of competition (i.e., on Assumption 9, not *API*).

For variations in investment policy, Fama and Arthur Laffer have shown that *API* holds given two or more noncolluding and price-taking firms with identical constant returns to scale technologies. *API* holds because one firm's investment decision has no effect on aggregate quantities supplied or demanded. There is no aggregate supply effect because investment expansions/contractions by one firm can and will be offset exactly by contractions/expansions by other firms. There is no demand effect because changes in investment by one firm have no effect on any individual's wealth (since with constant returns, all feasible investment levels have the same (zero) net firm value when evaluated at equilibrium prices). Importantly for our purposes, notice there is also unanimous indifference to the investment decision of a given firm (because no effect on prices and wealth implies no effect on consumption opportunities and utility).

Strict (unanimous) preference over investment decisions obtains given decreasing returns to scale as in the standard double-tangency proof of the *FST* (for example, see Fama-Miller, p. 70). But this proof assumes partial-equilibrium competition (i.e., Assumption 9, not *API*) because *API* does not hold in this case. With decreasing returns, one firm's investment decision has some (possibly small) effect on market-clearing prices. One firm will have some power to affect aggregate supplies since exactly offsetting supply responses by other firms will not generally be possible (see Fama-Laffer). It will also have some effect on aggregate demand due to different wealth (endowment) implications of different investment decisions. Thus, the standard decreasing returns assumption is consistent only with the partial-equilibrium unanimity theorem—that is, Theorem 2 cannot be extended to general equilibrium analysis via *API* in this case.

For variations in financing decisions, the conditions under which *API* holds also imply unanimous indifference to the decisions of a given firm.[12] Thus, for both investment

[12] *API* obtains in the standard Modigliani-Miller (M-M) world in which investment policy is assumed fixed and there are no corporate or personal taxes, no bankruptcy costs, and no agency costs. *API* obtains because 1) variations in capital structure by any one firm have no effect on aggregate supplies of state claim

and financing decisions, general equilibrium extensions of Theorem 2 via *API* lead to the uninteresting conclusion that individuals are unanimously indifferent to the decisions of a given firm. It follows that unanimity theorems which establish strict preference rather than unanimous indifference must rely on a partial-equilibrium notion of market competition.

IV. Summary

There is an extensive and growing literature which utilizes mathematically complex arguments to establish conditions under which corporate decisions are unanimously supported by firm owners. In this paper, I provided a simple and economically intuitive framework for understanding the critical conditions for unanimity. In particular, it was demonstrated that (in the absence of homogeneity assumptions on individuals and for quite general collective choice settings), the existence of unanimously supported decisions ultimately depends on the satisfaction of a simple opportunity set dominance criterion. I also formulated a general definition of market competition which includes as special cases the various notions of competition invoked in the unanimity literature and which implies that perceived value maximization is the unanimously supported corporate objective. The economic logic underlying the complex unanimity proofs in the literature was shown to be the same intuitive argument underlying the original Fisher Separation Theorem. I also distinguished unanimity theorems based on a partial-equilibrium notion of market competition (Assumptions 1–9) from those based on a general equilibrium notion of competition (Assumptions 1–9 and *API*). It was argued that unanimity theorems which establish strict preference rather than unanimous indifference must rely on a partial-equilibrium concept of competition. General equilibrium unanimity theorems require strong assumptions which imply the uninteresting conclusion that owners are unanimously indifferent to the decisions of a given firm.

Appendix

PROOF of Theorem 1:

Sufficiency follows directly from nonsatiation and the definition of global dominance. To establish necessity, assume that for each $\bar{z} \in Z$, there is at least one individual i such that $O^i[\mu^i[\bar{z}]|\pi^i]$ is not globally dominant. Let $\bar{y}^i$ denote i's optimal consumption bundle given decision $\bar{z}$. Since $O^i[\mu^i[\bar{z}]|\pi^i]$ is not globally dominant for i, there exists $\hat{y}^i \in O^i[\mu^i[\hat{z}]|\pi^i]$ with $\hat{z} \in Z$ and $\hat{z} \neq \bar{z}$ such that at least one element of $\hat{y}^i$ is strictly greater than the corresponding element of each $y^i \in O^i[\mu[\bar{z}]|\pi^i]$. Let i's utility function be L shaped through $\hat{y}$: $U^i[\hat{y}^i] > U^i[y^i]$ if any single element of y^i is less than the corresponding $\hat{y}^i$ element. Clearly, $\hat{y}$ is strictly preferred to all feasible $y^i \in O^i[\mu^i[\bar{z}]|\pi^i]$ including $\bar{y}^i$. Since $\hat{z}$ is feasible, $\bar{z}$ is not a solution to (1) for i and therefore is not unanimously supported. Since the argument applies to all $\bar{z} \in Z$, a unanimously supported collective decision does not exist and necessity is proved.

patterns (in the M-M world, financing decisions affect the packaging of claims, but not the social totals) and 2) distributive effects are assumed away either through Assumption 6 or by assuming that recapitalizations are nonexpropriative. *API* will not generally obtain with taxes, bankruptcy costs, or agency costs. Variation in one firm's financing decision will have some effect on aggregate supplies of state claims since taxes and leverage-related costs consume resources (i.e., social totals are not necessarily fixed) and since offsetting supply responses by other firms will not generally be possible. Distributive effects on demand will generally be present as well. This analysis explains why Stiglitz (1969, 1974) and Litzenberger-Sosin among many others have been able to prove general equilibrium leverage irrelevancy theorems for an M-M world while Alan Kraus and Litzenberger and many others have invoked partial-equilibrium notions of competition to prove that each firm has a unique strictly preferred debt-equity ratio in the presence of taxes and bankruptcy costs. This analysis also explains why Fama-Miller and Baron (1976a) invoked Assumption 9 to establish the partial-equilibrium result that the perceived value-maximizing investment policy (with decreasing returns technology) is unanimously supported, and used *API* to establish leverage irrelevancy *across* exchange equilibria in an M-M world.

REFERENCES

D. P. Baron, (1976a) "Default Risk and the Modigliani-Miller Theorem: A Synthesis," *Amer. Econ. Rev.*, Mar. 1976, *66*, 204–11.

______, (1976b) "Flexible Exchange Rates, Forward Markets, and the Level of Trade," *Amer. Econ. Rev.*, June 1976, *66*, 253–67.

______, "Investment Policy, Optimality, and the Mean-Variance Model," *J. Finance*, Mar. 1979, *34*, 206–32.

W. H. Beaver and J. S. Demski, "The Nature of Income Measurement," *Accounting Rev.*, Jan. 1979, *54*, 38–46.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

H. DeAngelo, "Three Essays in Financial Economics," unpublished doctoral dissertation, Univ. California-Los Angeles, Dec. 1977.

S. Ekern and R. Wilson, "On the Theory of the Firm in an Economy with Incomplete Markets," *Bell J. Econ.*, Spring 1974, *5*, 171–80.

Eugene F. Fama, "The Effects of a Firm's Investment and Financing Decisions on the Welfare of its Security Holders," *Amer. Econ. Rev.*, June 1978, *68*, 272–84.

______ **and A. Laffer,** "The Number of Firms and Competition," *Amer. Econ. Rev.*, Sept. 1972, *62*, 670–74.

______ **and Merton H. Miller,** *The Theory of Finance*, New York 1972.

Irving Fisher, *The Theory of Interest*, New York 1930.

S. J. Grossman and J. Stiglitz, "On Stockholder Unanimity in Making Production and Financial Decisions," *J. Finance*, May 1977, *32*, 389–402.

K. P. Hagen, "Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Mar. 1976, *66*, 199–203.

Jack Hirshleifer, *Investment, Interest, and Capital*, Englewood Cliffs 1970.

A. Kraus and R. Litzenberger, "A State Preference Model of Optimal Financial Leverage," *J. Finance*, Sept. 1973, *28*, 911–21.

H. E. Leland, "Capital Asset Markets, Production, and Optimality: A Synthesis," tech. rept. no. 115, IMSSS, Stanford Univ., Dec. 1973.

______, "Production Theory and the Stock Market," *Bell J. Econ.*, Spring 1974, *5*, 125–44.

______, "Quality Choice and Competition," *Amer. Econ. Rev.*, Mar. 1977, *67*, 127–37.

______, "Information, Managerial Choice, and Stockholder Unanimity," *Rev. Econ. Stud.*, Oct. 1978, *45*, 527–34.

R. Litzenberger and H. Sosin, "The Theory of Recapitalizations Under Incomplete Capital Markets and the Evidence of Dual Purpose Funds," *J. Finance*, Dec. 1977, *32*, 1433–55.

F. Milne, "Corporate Investment and Finance Theory in Competitive Equilibrium," *Econ. Rec.*, Dec. 1974, *50*, 511–33.

______, "Choice Over Asset Economies: Default Risk and Corporate Leverage," *J. Finan. Econ.*, June 1975, *2*, 165–85.

N. C. Nielsen, "The Investment Decision of the Firm Under Uncertainty and the Allocative Efficiency of Capital Markets," *J. Finance*, May 1976, *31*, 587–602.

G. Pye, "Present Values for Imperfect Capital Markets," *J. Bus., Univ. Chicago*, Jan. 1966, *39*, 45–51.

R. Radner, "A Note on Unanimity of Stockholders' Preferences Among Alternative Production Plans: A Reformulation of the Ekern-Wilson Model," *Bell J. Econ.*, Spring 1974, *5*, 181–84.

S. A. Ross, "Options and Efficiency," *Quart. J. Econ.*, Feb. 1976, *90*, 75–90.

M. E. Rubinstein, "Competition and Approximation," *Bell J. Econ.*, Spring 1978, *9*, 280–86.

J. E. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

______, "On the Irrelevance of Corporate Financial Policy," *Amer. Econ. Rev.*, Dec. 1974, *64*, 851–66.

The Homogenization of Heterogeneous Inputs

By James M. Buchanan and Robert D. Tollison*

> The absurdity of treating land-use as a homogeneous magnitude has been commented upon above; regarding labor, the fallacy has been pointed out often enough — quite typically by writers who go ahead to discuss "wages in general," as if the concept had meaning.
>
> *Frank H. Knight*, p. 76

This paper examines the allocative effects of institutional constraints that require purchasers to treat *heterogeneous* inputs as *homogeneous* for input pricing. Our attention was drawn to this problem by a comparative analysis of salary policies in academic institutions, but further exploration suggests widespread applicability and relevance in such areas as 1) the differential allocative effects of craft and industrial unionization of the labor force; 2) the allocative effects of "equal pay for equal work," either as an institutionally adopted "principle of justice" in compensation, or as a result of legal prohibitions on discrimination in payment; and 3) the channelization of profit- or rent-seeking activities, and the subsequent processes through which profits or rents are dissipated. We shall not attempt to discuss these, and possibly many other, applications of our analysis, except for purposes of illustrating the argument. We shall concentrate on the elementary analysis.

Orthodox price theory contains a well-developed analysis for the pricing of homogeneous inputs, even when at another level it is acknowledged that inputs may in many cases be unique. Standard procedure simply imposes homogeneity by abstraction, a step that has often been the focus of criticism. This theory of input pricing is not wholly consistent with the theory of the organization of production, which is usually considered to involve the combination of separate inputs synergistically, inputs that are by implication heterogeneous rather than homogeneous. Indeed, the role of the entrepreneur is widely interpreted to be that of sensing potential profits from new input *combinations*. Our analysis allows both the organization of production and the entrepreneurial role to be more readily brought within a framework analogous to that of the orthodox theory of input pricing.

I. Heterogeneity and Homogeneity

Heterogeneity and homogeneity may, of course, be defined in many different ways, and the problem is to define and to use these terms appropriately for the analytical purposes at hand. One source of common confusion lies in economists' proclivity to think of inputs (and outputs) in physical dimensions rather than in terms of valuation. Even within the valuation rubric, however, it is necessary to distinguish categorically between *internal* and *external* valuation, between the two separate sides of choice, so to speak.

In the internal evaluation of separate input units, the prospective purchaser or demander could classify units by homogeneous bundles, with homogeneity defined in terms of predictions of potential contributions to product value. Such a calculus would, presumably, be embodied in the ultimate demand for input units. Our concern here is not with the demander's *internal* evaluation of input units, save insofar as this enters the demand function. Rather our direct concern is with the *external* evaluation of inputs, which are presented to the potential purchaser as market-determined input prices. These prices reflect the opportunity costs for such input units, as perceived by the owners of the resource inputs themselves. In our context, the owner of an

*Center for Study of Public Choice, Virginia Polytechnic Institute and State University. We are grateful to Thomas Borcherding, Geoffrey Brennan, Roger Congleton, Roger Faith, Robert Mackay, Robert McCormick, Dennis Mueller, and Gordon Tullock for helpful comments on earlier drafts of this paper. The usual caveat applies.

input foregoes value (pecuniary and nonpecuniary), measured by the amount that the unit may command in alternative uses to that which involves the sale to the prospective purchaser. The difference between the opportunity cost perceived by the owner of a resource unit and the actual price or wage received for the unit in a transaction is economic rent in the usual definition. This rent will figure prominently in the analysis that follows.

In terms of external evaluation, homogeneity and heterogeneity must be defined by the prices that the prospective purchaser or demander faces. If input A is available for purchase at the same price (wage) as input B, these two units are *homogeneous* in external value terms, regardless of their possibly differing anticipated contributions to product value for the purchaser (and hence their possible internal heterogeneity), and regardless of any similarity or difference in physically observable attributes. If, on the other hand, input unit C can be purchased at a price (wage) that is different from that required to purchase input unit D, these two units are *heterogeneous* in external value terms, regardless of their possible differences or similarities, either in internal evaluation or in physically measurable attributes.

II. A Model of Heterogeneous Input Combination

We now introduce a model that includes a profit-maximizing firm (or entrepreneur) which senses an opportunity for putting together a particular combination of input units. This combination includes capital and management inputs that are purchased competitively in ordinary markets. But it also includes a bundle of "labor inputs." By assumption, these latter units are heterogeneous in the external evaluation sense defined above. For simplicity in exposition, we assume that the prospective input combination, the "labor bundle," includes only one unit from each type or kind. The analysis could, of course, be extended to apply to the combination of several homogeneous units from several heterogeneous groups. Each unit is available at a price that is determined in a competitive market setting, and the firm has no influence over the price that it pays for any particular resource unit.

In order to simplify the exposition initially, we assume that the inputs are internally valued at the same level. The entrepreneur imputes an equal anticipated product value to each of the units in the "labor input bundle," despite the fact that he also recognizes that the separate units "do different things" and, hence, are less than perfectly substitutable one for another. Our conception of production here is that of a synergistic, team production process. Part of the process includes other inputs, such as capital, which, as noted above, are purchased competitively, and part includes the separate, heterogeneous labor inputs. All of these inputs are combined by the entrepreneur to produce an output expressed in terms of a market value. The entrepreneur must decide what market value of output to produce, and he must hire and combine the units of productive inputs to produce this output.

Armen Alchian and Harold Demsetz have analyzed the problem of monitoring and disciplining inputs in a team production process. Our concern is with the analysis of the hiring of the heterogeneous labor inputs used in such processes. The production process that we analyze does not embody a fixed-coefficient production function. Within limits, entrepreneurs can substitute among categories of labor inputs. A contractor can build a house with wooden or plaster walls, with different labor inputs as well as capital, with carpenters or plasterers. If plasterers have a lower competitive supply price than carpenters, the contractor can build more houses with plaster walls. This substitution among heterogeneous labor inputs will show up in the market value of the contractor's housing output. But there is a point in this process of substitution beyond which the contractor-entrepreneur cannot go; he cannot build his houses completely out of plaster. We shall return to discuss the concept of team production with heterogeneous inputs at a later stage in the paper. Suffice it to say here that the conception of team production that we envisage for purposes of analysis is quite different than that used for the conventional analysis of the pricing of inputs, in which different units of an input are readily substitutable for one another.

In the model as postulated, we can now build up a "supply schedule" for the separate "labor" inputs that the entrepreneur will face, as depicted in Figure 1. Note that this schedule is not the same as that which is familiar in orthodox analysis, where the input units measured on the abscissa are, by definition, homogeneous in some "physical" sense. In the schedule of Figure 1, input units are arrayed along the abscissa in the order of the ascending market prices for hire. The marginal valuation or demand curve is drawn in the usual manner. The size of the input combination, and hence the rate of output production, is assumed to be variable, even though inputs undertake differing tasks. The entrepreneur will extend his purchases to Q, which is the efficient rate of input use, given his evaluations. At Q, the last unit of input hired will, of course, be paid at the E_1 rate. All other units will be paid in accordance with their market-determined prices, all lower than E_1.

We should stress the nature of the upward-sloping "supply curve" S in Figure 1. This curve differs from the standard supply relationship in the heterogeneity of units measured. The curve is, however, a genuine marginal labor input cost curve for the firm, and the equation of marginal input cost with the estimated value of product at E_1 mirrors the standard efficiency requirement. The profit-maximizing firm or entrepreneur will combine units of labor input (Q in number) out to E_1, however, only because he faces S as the marginal cost curve. This amounts to saying that he must be allowed to purchase each of the input units at the market price established externally *for each unit of input*.

This purchase "policy" appears to duplicate that which might be adopted by a monopsonist, who is somehow able to discriminate perfectly among separate suppliers of homogeneous inputs. The perfectly discriminating monopsonist will also be led

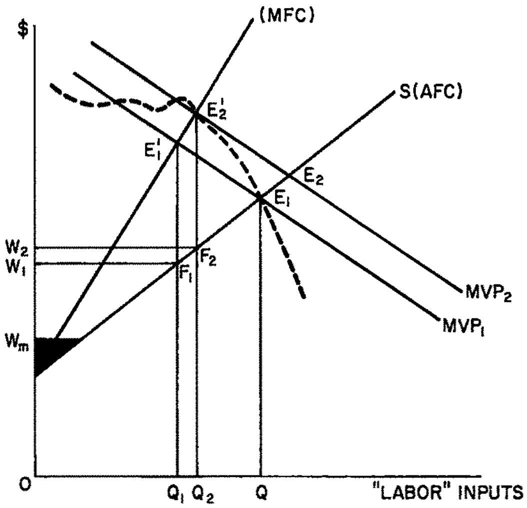


FIGURE 1

to the efficient quantity of resource use.[1] What is interesting in our model, however, is the absence of *any* market power on the part of the purchasing firm. We can assume full competition in the firm's product market, along with competition in each one of the input markets that it faces. Efficiency emerges from the profit-seeking behavior of all participants in the interaction. This result is, of course, in line with the neoclassical orthodoxy.[2]

[1] Buchanan, in a discussion of military manpower, observed that "the fact that the supply curve slopes upward indicates differential productivity in alternative employments despite the homogeneity of units in producing military services" (p. 90).

[2] To our knowledge, aside from Buchanan's precursory discussion, the standard theory of input purchase and pricing contains little treatment of the particular implications of input heterogeneity emphasized in this paper. As the introductory quotation suggests, Knight often criticized the economic theory of input pricing for failing to come to grips with the implications of input heterogeneity, but he never expanded upon this basic criticism. Martin Bronfenbrenner's early application of monopsony theory to input pricing offers a partial exception. The emphasis in the standard approach is typically on the purchase and pricing of homogeneous inputs under various market conditions. See, for example, the discussion in Milton Friedman, pp. 176–93.

III. Institutional Homogenization

We now want to apply our analysis to examine the effects of an input pricing policy that requires equalization of payments (wages) for some or for all the heterogeneous units that are purchased by a firm.[3] In long-term equilibrium, no firm earns profits. From this condition, it follows that *any* increase in costs must generate losses, and, in consequence, some exit of firms from the competitive industry. In the context of heterogeneous input purchases, any requirement that firms pay above-opportunity-cost prices or wages increases the cost of the firm's output, generates losses, and ultimately reduces the number of firms in the affected industry. Industry output declines, and product price goes up. There will be some welfare loss approximated by a Harberger-type triangle underneath the demand curve and above the cost schedule for the industry's product over the relevant adjustment range. This welfare loss will be directly related to the size of the *rents* created by the homogenization in pricing.

A truncation of the lower bound to input prices (wages) may force homogenization over some range of inputs without modifying the apparent equilibrium of the firm, as depicted at E_1. For example, suppose that a minimum-wage law requires all labor to be paid a wage of W_m, generating rents as shown by the darkened triangle for the lowest opportunity cost workers. The firm will still face the marginal input cost curve S beyond the height W_m, and will attain equilibrium at E_1. But any firm at E_1 under these conditions will be one among a smaller number of firms than would have been the case without the minimum-wage restriction. Any increase in the minimum-wage restriction will increase net welfare losses, as well as the net transfer of rents to the marginal input owners who remain employed.

Our primary interest, however, lies in determining the effects of a general requirement that prices (wages) be uniform over *all* units of inputs that are hired or purchased. Suppose that such a policy comes to be enforced by a law against differentiation in compensation among labor inputs. If this policy should be forced on one firm in isolation, it could not survive under the competitive conditions postulated. We seek to examine the effects of such a policy that is enforced for all firms in the product group, all of which have comparable production functions.

Initially, we assume that the individual firm is allowed to select the level of the uniform input price it is to pay, along with the number of input units it will purchase at that price. The S-type curve of Figure 1 will no longer represent a marginal input cost curve to the firm under these conditions. This curve will now reflect a schedule of *average* input costs, the familiar construction from orthodox monopsony theory. At any point along such a curve, costs will be higher to the firm than under opportunity cost pricing by the amount measured by the triangle bounded by the S curve, the uniform price, and the ordinate. Faced with the uniform payment requirement, firms initially operating at position E_1 will seek short-term adjustment by shifting to E_1', determined by the intersection of the MVP_1 curve and the new curve of marginal input cost. The short-run price will be W_1, with Q_1 units of input employed.

Even with these short-term adjustments, however, firms will make losses because of the increased costs in rent transfers, and some firms will leave the industry. As this long-term adjustment takes place, total industry output falls further, and demand price will rise, which will in turn increase a firm's marginal value product estimates for input units. For a firm that remains in the industry, a new short-term equilibrium will be attained at some position like E_2', with an input purchase rate of Q_2. The firm will also be in long-term equilibrium with zero profits or losses.

The input price or wage W_2 defines that wage or price for which the uniform compensation requirement can be met with

[3] In the context of an application of conventional monopsony theory, Thomas Borcherding suggests the importance of assumptions about equal wages for identically defined jobs.

minimum welfare losses. By construction, any level below W_2 will encourage the firm to expand employment by paying differentially higher prices for extramarginal units. Any level above W_2 will satisfy uniformity only at increasing welfare costs.[4]

By looking at E_1 and E_1' (or at E_2 and E_2'), the homogenization of the heterogeneous inputs in pricing seems to have the effect of making the firm in a fully competitive setting, in both product and factor markets, behave as if it were a genuinely nondiscriminating monopsonist. In this context, note that the total welfare losses cannot be measured by the triangles $(E_1E_1'F_1)$ or $(E_2E_2'F_2)$, multiplied by the number of firms. Such a measure offers only a short-term approximation to these costs. If we think of n_1 firms in the initial competitive equilibrium depicted at E_1, welfare costs in the short term are approximated by $n_1(E_1E_1'F_1)$. As firms leave the industry, however, total welfare costs are reduced, and if n_2 firms remain in the industry in some new long-term equilibrium, welfare costs of the uniformity requirement are then minimally valued at $n_2(E_2E_2'F_2)$. In each measure, however, the number of firms is fixed, whereas any accurate measure of welfare costs must embody long-term adjustments in industry output and price, as evaluated by final consumers of the industry product.

To this point, we have explicitly assumed that the inputs in the labor bundle, defined as heterogeneous in terms of their supply prices to the firm, are homogeneous in terms of their anticipated internal productivity to the firm. The analysis fully applies to all bundles of possible inputs that can be grouped by internal valuations in this manner. The apparent restrictiveness of the internal homogeneity assumption may be relaxed, however, without destroying our results, provided we are careful to specify the side conditions that must be met. For any set of input units that might be assembled, the set of externally determined input prices provides a natural basis for arraying the separate units along the abscissa, and, hence, for the construction of the marginal cost or supply schedule. Unless we assume homogeneity among the units in terms of anticipated internal productivity, however, things become considerably more complex in deriving the demand or marginal evaluation schedule. Once we make such an assumption, we are able, of course, to devise a marginal valuation schedule that is precisely analogous to that in orthodox input pricing theory, one in which the law of diminishing returns and the demand schedule for the product determine the single valued relation between input usage and value. But in the absence of internal homogeneity, why should all potential inputs in team production be expected to produce equal increments to output value?

The evaluation of a specific unit of input, described by its place in the prescribed array, will reflect some combination of two quite separate elements. First, units placed first in the array will tend to generate higher increments to product value by the ordinary law of diminishing returns, considering organization and other inputs as the fixed factors. Secondly, however, units may, *ceteris paribus*, be estimated to contribute differentially to product value. The combination of these elements may be such as to introduce nonconvexities in the demand or evaluation schedule of the firm for the set of inputs as arrayed. Stability requires that the demand or marginal evaluation schedule cut the marginal cost schedule from above, but, beyond this, little can be said about particular shapes in the most general setting. If, however, we assume that the schedule is convex from below, the general conclusions reached above about the introduction of uniformity in compensation hold.

[4] Our construction differs from that of familiar monopsony theory in one important respect that deserves mention at this point. In the orthodox analysis of the nondiscriminating monopsony firm, the externally imposed requirement that the firm pay some wage above that average factor price it would optimally select can result in an expansion in the employment of labor by that firm. Since, by construction, labor's marginal product in the firm exceeds that elsewhere in the economy, the increased employment is welfare improving. By contrast, in our construction, *any* increase in rents increases costs in the competitively organized industry, reduces industry output, and with normal input relationships, must *reduce* employment. The *curiosus* of orthodox monopsony theory is not applicable here.

Even if we allow for nonconvexities, so long as the evaluation curve does not, at any point over the inframarginal range, fall below the profit-maximizing uniform rate of payment (W_2 in Figure 1), the conclusions continue to apply. One such schedule is depicted by the dotted curve in Figure 1. If, however, the schedule should be sufficiently "misbehaved," forced equalization of payment for inputs would result in a complete shift in the team's composition. Units of input at differing places in the initial array might be dropped, and a new array might be formed, once again in terms of external prices in the absence of forced equalization. Given such a new array, our model would again fully apply, with units of inputs at the lower end of the array receiving relatively higher rents from equalization policy.

IV. Team Production with Heterogeneous Inputs

Consider, as an example, an entrepreneur who is organizing a construction team. Suppose that the potential members of this team include, in ascending order of competitively set wage rates: a laborer, an excavator, a painter, a carpenter, a plasterer, a mason, a tile setter, an electrician, an insulator, a plumber, and a cabinet maker. Suppose that, at the prevailing competitive wage rates established separately in each occupational or skill group, all eleven of the potential members are employed. (We could, of course, allow for differing numbers from each group.) By the terms of the previous analysis, if the entrepreneur's projections are accurate, the outcome is efficient, locally and globally.

Now suppose that someone petitions for unionization of the whole team. An election is held, and suppose that at least six members of the group vote to unionize. Subsequently, suppose that the union demands, and gets, an agreement for uniformity in compensation over all members of the team, but that the firm is allowed to set the uniform wage to be paid and the number of workers to be hired. It is clear that those workers whose opportunity costs are relatively low and who remain employed will gain rents in the process. The workers with relatively high opportunity costs may secure positive but lower rents if they remain employed on the team, and they will lose only the transaction costs involved in a shift to alternative employments if they are not retained as members of the team. The team's organizer, the entrepreneur, will find his costs per unit of output increased, and will undergo short-term losses until adjustments to the new cost schedule are made. Consumers of the team's product will lose surplus in the amount of the rents gained by the inframarginal inputs, plus the excess social loss measured by the value of output lost over its true marginal cost.

If the union is allowed to set the uniform wage, to apply over all of the labor inputs hired, this wage need not be that which will be profit maximizing for the firm, given the uniformity requirement. The union-set wage may be above W_2 in Figure 1. In this setting, the firms that remain in the industry after all adjustments may each employ more than Q_2 from the bundle of heterogeneous inputs, but there will be fewer firms in the industry than under the firm-set uniform wage plan. And, since total rents are higher, total welfare losses will exceed those present when firms are allowed to select their preferred uniform levels of compensation.

Implications for the predicted effects of differing forms of unionization emerge clearly from the analysis. Suppose that, in an alternative scenario to that sketched out above, carpenters unionize across all employers and set standard rates. These rates will tend to lie above competitive levels, and the employment of carpenters will be adversely affected. But, per head, carpenters will secure higher returns than in the absence of such unionization. Generalized unionization and standard wage setting within each of the crafts or skills, within each of the heterogeneous input sets separately, need not, however, result in an elimination of the differentials in compensation rates among the different crafts, as faced by the employing entrepreneur or firm. The organizer of team production may still confront an array of input prices for the heterogeneous inputs considered as potential members of the team, and given such an array, he may be led to some admittedly second or third best optimum rate of input usage.

The familiar differences between craft and industrial unionization emerge directly from the analysis. Relatively low-skilled workers will tend to secure differentially higher gains from unionization and wage standardization across some or all heterogeneous inputs employed by a given firm, or by all firms producing a given product, provided that they expect to remain employed after unionization. Hence, we should predict that pressures for such unionization, and for the tendency toward standardization of input prices that would follow, will be greater from those workers whose skill category may be relatively low, but whose contribution to team production is estimated to be relatively higher than their higher-skilled cohorts. (Skills are defined here implicitly by relative opportunity costs in competitive labor markets, which would presumably reflect some equilibrium of supply and demand within separate input groups.) By comparison with the low-skilled workers, those in the relatively high-skilled "professions" or "occupations" may secure little if any gain from cross-input wage standardization. Members of these groups may expect to secure net rents from successful unionization across all members of each narrowly defined profession, which would presumably succeed in increasing wages above competitive levels.

As we noted, our attention was drawn to this problem by comparing alternative salary policies in academic institutions, which seem to offer excellent real world examples of the heterogeneous input phenomena. The college or the university can be modelled as a team, one that combines heterogeneous inputs to produce some "education-research" services. Potential contributors to the team may include professors (instructors) from various academic disciplines: English literature, history, philosophy, sociology, political science, psychology, biology, chemistry, physics, economics, engineering, business, law, medicine, and so forth. As arrayed here, competitive salary levels in each of the labeled disciplines stand in some roughly ascending order. Pressures on the university toward uniformity in salary scales would have the effects traced out in the basic model of Sections II and III.[5] Faced with requirements for equalization of salaries across heterogeneous disciplines, universities will tend to drop (or fail to add) high-salary disciplines to the "academic team" that is organized to produce academic output. It also follows that demand shifts will generate larger shifts in the employment of team members from high-salary than from low-salary disciplines.

V. Equal Pay for Equal Work

The examples discussed suggest that the separate inputs combined in team production are indeed heterogeneous, and not only in the narrowly defined sense introduced earlier in the paper. Painters *are* different from carpenters; historians *are* different from economists. This natural (and nonscientific) classification of inputs or factors in terms of visible, and apparently observable, physical attributes is helpful in gaining some acceptance of the analytical conclusion that any artificially contrived and enforced homogenization for pricing purposes must generate losses in allocative efficiency. In a comparable sense, however, the same natural tendency to classify inputs by physically measurable attributes or dimensions creates difficulties in securing the acceptance of the extension of the identical analytical conclusion in a setting where inputs may be heterogeneous in the external valuation sense, but where they may "seem" homogeneous in physical characteristics, and where they may also be homogeneous in internally estimated contributions to product value.

To imply that worker C, who seems identical in physical appearance to worker D, and who, furthermore, is observed to produce precisely the same physical output as

[5] Academic institutions tend to be nonproprietary rather than profit maximizing. This difference between these institutions and private firms make universities more vulnerable to pressures for standardization. The welfare effects of standardization under non-profit-maximizing conditions are similar, but not identical, to those derived above for profit-maximizing firms. Geoffrey Brennan and Tollison offer a more detailed examination of the academic institution's behavior and its allocative effects.

worker D (both may, for illustration, be seen to operate similar machines that produce six widgets per hour), is *different* for wage payment purposes from worker D because C's opportunity cost (his external earning capacity) is different, runs immediately afoul of commonly held ethical criteria for "justice." Such criteria dictate that such apparently equal workers *should* secure equal payment. It becomes much more difficult in such a setting to convince skeptics that justice, legally secured, is gained at the expense of allocative efficiency.

In our analysis, even if separate units of input are physically identical, and even if each unit adds the same amount to product value for the firm, these inputs are *heterogeneous* if their opportunity costs differ. Any artificially enforced policy that requires firms to follow an "equal pay for equal work" guideline, therefore, must be allocatively inefficient. This is not to suggest, of course, that equal pay for equal work as a result or end state may not emerge from the normal workings of the market process under many, perhaps most, circumstances, and particularly those circumstances that are characterized by freedom of entry and exit by firms in separate product lines and by workers in separate employments. However, as a criterion for some legally enforced wage setting, equal pay for equal work cannot fail to have the effects suggested.

In terms of the efficiency norm, "equal work" is inappropriate as the criterion for "equal pay." This criterion should be "equal cost," which is, of course, that which the market will tend to meet if its operations are unimpeded. The equal work criterion will be satisfied only if there exist no equalizing differences present that will introduce differences among opportunity costs at the margins of adjustment. And, of course, there is no way of determining whether or not such differences exist, other than that of allowing the market itself to work freely.

Issues of "discrimination" easily arise here. Is it not "discriminatory" to allow employers to pay differing wage rates to workers who produce equal measured products? Again, reference to our earlier example will prove useful in clearing out tightly packed emotional cobwebs. No one would think of discrimination if he should observe a contractor paying a plumber a higher hourly wage rate than a painter. Why, then, should discrimination seem to be present when worker C is paid a lower wage rate than the equally productive worker D, if the employer can hire C at a lower opportunity cost?

Suppose that, upon examination, worker C should be found to have a strong locational preference for employment in place E where the employment is, and that he will indeed accept employment at E at some wage rate considerably lower than that required to hire worker D. Equalization of wage rates, at the level that D requires, will amount to paying C an economic rent over and above his true opportunity cost, and the necessity of paying this unnecessary rent will force the employer to reduce the rate of input purchase (and of output production) below that which is socially efficient.[6]

VI. Input Heterogeneity in a Generalized Setting

In this section we want to examine the possible relevance of the phenomena we have analyzed in a more general setting, and specifically to examine how the results might be modified in the long run. To what extent is the input heterogeneity that is central to our analysis a short-term or disequilibrium phenomenon that will tend to disappear as resources shift into and out of alternative categories? To what extent does the input composition of team production change as adjustments might be made to the differing opportunity costs of inputs? To what extent

[6] In the specific setting discussed here, the single employer would be interested in hiring worker D at the differentially higher wage required, only if he exhausts the supply of workers like C, given an assumption that initially the C and D workers are interchangeable. In this setting, therefore, the firm is a monopsonist of the traditional variety. Our analysis suggests only that the discriminatory input pricing that the monopsony firm would naturally follow is required for allocative efficiency. The monopsonist need not, of course, retain the "rents" that might otherwise be gained by input owners. Competition in the product market may allow the monopsonist to survive only if he is able to discriminate in the purchase of inputs.

will consumers tend to shift purchases toward products and services that embody relatively high proportions of inputs that are priced at relatively low opportunity costs?

For all practical intent, our analysis has potential policy relevance only for personal or human inputs into the productive process. Land and capital will normally be priced by the play of ordinary market forces, and any proposed forced homogenization of heterogeneous units would appear, and be, absurd. Consider, then, labor input alone, and look at the possible substitutability among heterogeneous input categories within the inclusive labor services rubric, that is, substitutability among differing occupational, skill, and professional groups. As noted previously, the competitive wage level that will emerge for any given group of homogeneous units within a single category will depend on some amalgam of both demand and supply elements.

If, from some preexisting or postulated equilibrium wage level relationship between two separate categories, say between carpenters and plumbers, there occurs an exogenous demand shift that modifies the relationship, marginal adjustments will occur. Some workers will undergo retraining to shift from the relatively lowered to the relatively increased wage occupation. New entrants will tend to enter more rapidly in the latter occupational category than the former. These long-term adjustments will tend to return the relationship between wage levels in the two occupations to their previously existing equilibrium. But there is nothing in such long-term adjustments that insures wage rates in the two categories will achieve ultimate equality. "Equalizing differences" may remain as between occupational categories, even with long-term adjustment, differences that may be due to differences in the intrinsic attractiveness or unattractiveness of employments, or due to natural limits on the supply of specific talents among persons. So long as such equalizing differences are acknowledged to exist in the long-term equilibrium adjustments in the supply of labor to the different occupational and professional groupings, our analysis remains unaffected.

A different sort of substitutability may take place, however, within the production function of the organizer of production. To the extent that the heterogeneous inputs are substitutable one for the other in the production function, the profit maximizer will, of course, prefer to purchase those units that can be hired at the lowest opportunity cost. This activity on the part of firms will tend to increase demand for the low opportunity cost categories, and to bring wages or prices toward equality with those for other inputs. This equalization process will continue so long as the substitution possibilities are efficient. As we stressed at the outset of our analysis, however, to the extent that there remain genuine advantages to team production, such input substitution is limited, even with long-term adjustments. Much the same conclusion holds for the long-term substitutability among final product items by ultimate consumers. There are finite limits beyond which substitution motivated by price differences becomes inefficient.

If input heterogeneity arises from differences in opportunity costs that embody no equalizing differences, and there are no advantages in team production that require some combination of differing talents or skills, "unequal pay" for "equal work" becomes a phenomenon to be observed only in short-term or disequilibrium situations. In the literal definitional sense, unequal pay for equal work is inconsistent with the conditions required for long-term competitive equilibrium. As noted earlier, equal pay for equal work tends to emerge, where it is relevant, from the working of market forces. Even in this context, however, any premature and forced homogenization of pricing for the initially heterogeneous units of input will prevent the allocative adjustments that are required to attain the end-state equalization in compensation without coercion.

Consider as an example a group of immigrant workers who initially have very low opportunity costs and who will accept employment at relatively low wages. Firms that hire workers will be observed to make differentially higher profits. Demands for the services of such workers will increase, and wages will tend toward equality with those of other groups with comparable skills. Suppose, by comparison, that all employers are forced to pay equal wages from the outset. There will now be no profit opportunities in hiring the immigrants. The rate at which members of this group are absorbed into the labor force will be retarded, and the economy will suffer a welfare loss due to resource misallocation over a longer time period.

VII. Conclusions

Orthodox theory assumes that inputs are readily classifiable into groups that internally contain units that are indistinguishable, one from another, both in opportunity costs or in contributions to value. The theory of pricing is then applied to each type or class of input separately considered. Prices are set in accordance with the values of marginal product, but there is no difference between what is sometimes called the "internal" marginal product and the "external" marginal product. No problem arises concerning the definition of heterogeneity or homogeneity among input units. As noted, the relationship among the prices for input units drawn from different classes and combined by the firm has not been analyzed directly to our knowledge, and little or no attention has been paid to the allocative effects of forced homogenization in pricing for heterogeneous input units. As our applications suggest, however, this sort of homogenization is widely observed in real world economies, and notably in connection with labor inputs.

There are two separate but related strands of analysis in neoclassical theory that are analogous to the analysis in this paper. Constructions that seem almost identical to our own are to be found in the orthodox monopsony theory. Perhaps the most interesting feature of our model lies in the demonstration that our results emerge even if the firm operates within a fully competitive structure, with no influence on either the price of its product or the price of any of the inputs that it purchases. In this context, the firm that organizes team production among units of heterogeneous inputs behaves analogously to a perfectly discriminating monopsonist. It attains allocative efficiency in input usage (and output production) because it is able to secure the heterogeneous input units at their differing opportunity costs, measured in the competitively determined set of input prices.

The upward-sloping "supply curve" for heterogeneous inputs may be confronted by firms that remain extremely small relative to the market for either the product or for any one of the inputs purchased. The firm need not be large relative to the market for inputs for the upward-sloping cost curve to emerge, as would be the case in the standard derivation of a monopsony position. The forced homogenization has the effect of making the fully competitive firm act as if it becomes a nondiscriminating monopsonist. In both cases, welfare losses emerge. However, the ultimate policy implications are quite different in the two cases. The welfare losses of nondiscriminating monopsony can be eliminated by the introduction of competition in the markets for inputs. The welfare losses from forced input homogenization can also be eliminated by the "introduction of competition," but there it takes the form of allowing ordinary markets to operate without interference.

The second strand of analysis that is analogous, although less directly, to our analysis is found in the famous discussion of rising supply price that occupied the best minds of the profession in the 1920's and 1930's, with the debated issues finally resolved in Joan Robinson's paper. Marshall and Pigou had argued that competition in industries characterized by a rising supply price tends to produce an inefficiently large output. The opponents of this view finally proved that competitive organization generated the optimal output, that rents did not constitute social losses, and that any restriction below the competitive output levels would guarantee welfare losses. In the models of rising supply price under competition, however, individual firms face parametric input prices defined over bundles of separately considered homogeneous input units. Firms do not face rising supply prices for inputs arrayed in terms of competitively established opportunity costs. While our results are broadly consistent with the conclusions reached in this great neoclassical debate, they clarify an ambiguity in demonstrating that competitive output is efficient even when the competitive firm itself faces upward-sloping input cost schedules.

REFERENCES

A. A. Alchian and H. Demsetz, "Production, Information Costs, and Economic Organization," *Amer. Econ. Rev.*, Dec. 1972, *62*, 777–95.

T. E. Borcherding, "A Neglected Social Cost of a Voluntary Military," *Amer. Econ. Rev.*, Mar. 1971, *61*, 195–96.

H. G. Brennan and R. D. Tollison, "Rent Seeking in Academia," in James M. Buchanan, Robert D. Tollison, and Gordon Tullock, eds., *Towards a Theory of the Rent-Seeking Society*, College Station 1980.

M. Bronfenbrenner, "Potential Monopsony in Labor Markets," *Ind. Labor Rel. Rev.*, July 1956, *9*, 577–88.

James M. Buchanan, *Cost and Choice*, Chicago 1969.

Milton Friedman, *Price Theory*, Chicago 1976.

F. H. Knight, "The Ricardian Theory of Production and Distribution," in his *On the History and Method of Economics*, Chicago 1956.

J. Robinson, "Rising Supply Price," in George J. Stigler and Kenneth E. Boulding, eds., *Readings in Price Theory*, Homewood 1952.

Bank Regulation and Macro-Economic Stability

By ANTHONY M. SANTOMERO AND JEREMY J. SIEGEL*

Recent congressional legislation allowing financial institutions to pay interest on deposits, as well as allowing the Federal Reserve to pay interest on bank reserves, has again raised the question of the effects of such proposals upon the stability of the macro economy.[1] In the early 1960's when these issues were previously discussed, the work of William Brainard and James Tobin and Brainard were significant contributions to the area. These authors determined the conditions under which changes in banking regulation would be expansionary or contractionary for the economy as a whole. Brainard correctly noted that the differential response to exogenous variables induced by regulatory changes in the structure of the banking system cannot be used to assess the desirability of such a shift. In a deterministic environment, whether changes in the regulatory environment are expansionary or contractionary is totally irrelevant. If the system is nonstochastic, the monetary authority can often costlessly alter the size of the policy variables so as to compensate for any change in the response of the target variables to autonomous shifts.

The responsiveness of the system to exogenous shocks becomes most important in a stochastic environment. As William Poole has shown in an *IS-LM* framework, the failure of the monetary authority to assess accurately the current values of all endogenous variables indicates that alterations in the financial structure may have important implications for the success of the central bank stabilization policy.

This paper considers the effect of recently enacted legislation contained in the Depository Institutions Deregulation and Monetary Control Act of 1980 as well as other proposed policy recommendations on the monetary authority's ability to control its target variables in a stochastic environment. It indicates that some of these changes would categorically worsen stochastic control, while other seemingly similar proposals would have the opposite effect. The optimal movement in the structure will depend, as in Poole's paper, upon the assumed characteristics of the stochastic nature of the system.

It should be noted that while the motivating issue of the present study is the recently enacted bank legislation in the United States and the resultant change in banking regulation, the techniques proposed to evaluate the changes are most general. They can be employed to address a whole range of structural changes in the financial system in an analogous manner to the present work. The contribution here should be viewed as the development of techniques to evaluate other more general debates on the optimality of financial structure.

The paper will proceed as follows: Section I will outline the general equilibrium model employed in the study, as well as its stochastic behavior. Section II outlines the proposed regulatory changes and integrates them into the financial market model developed in the previous section. Here the impact of regulatory change on the variance of the target variable, the price level, will be

*Professor and associate professor of finance, respectively, The Wharton School, University of Pennsylvania. We wish to thank George Borts for a number of valuable comments on an earlier draft. This paper was begun while Santomero was visiting professor at Ecole Superieure des Sciences Economiques et Commerciales, France, and at the European Institute for Advanced Study in Management, Brussels.

[1] The Depository Institutions Deregulation and Monetary Control Act of 1980 was passed into law as Public Law 96-221. See below for a review of its substance. The literature is replete with proposals for the type of regulatory change analyzed here. To list the more notable recent contributions, a discussion of interest on reserves can be found in the study by Ira Kaminow, "Why Not Pay Interest on Member Bank Reserves?" Proposals and analysis of interest on deposits are detailed in the Board of Governors study, "The Impact of the Payment of Interest on Demand Deposits," and R. Alton Gilbert, "Effects of Interest on Demand Deposits: Implications of Compensating Balances."

considered explicitly. Section III introduces the real sector into the model and evaluates the effect of financial change upon the system's response to real sector disturbances that are stochastic in nature. Section IV integrates the real and financial sectors, demonstrating which policy shifts are appropriate under differing origins of the stochastic disturbances.

I. The Model of Financial Equilibrium

A. *A Description of the Static Model*

The basic framework employed in the analysis is similar to the general equilibrium framework of the financial markets first presented by Brainard and Tobin. Three sectors exist in the economy: the household or public sector h; the banking or financial institution sector b; and the firm sector f. Each sector's demands and supplies satisfy balance sheet constraints and substitution properties that are characterized by gross substitutability and normality.[2] There are four assets in the economy: high-powered or base money H, which also serves as currency; deposits at financial institutions D; bonds issued by the firm sector B; and equity E.

Within the four markets each sector is constrained by a budget constraint limiting asset purchases. For the household, real asset demand is constrained by wealth, so that

$$H_h^d+D_h^d+B_h^d+E_h^d=W\equiv K+H/P$$

where wealth has been defined as total firm capital plus high-powered money in real terms.[3] The financial institution sector, as an internal financial entity, must sum to zero in real value so that its budget constraint may be written as $H_b^d+B_b^d=D_b^s$ where superscript s indicates supply and, following *U.S.* regulation, these institutions are excluded from owning equity. Finally the firm sector obtains funds from debt and equity to maintain its fixed capital and holds no high-powered money nor deposits. Therefore, its constraint is $B_f^s+E_f^s=K$.

Formally, the model containing these constraints and the conditions of gross substitution and normal goods may be constructed as the following four-equation system of equilibrium conditions:

$$(1)\quad H_h^d\left(\overset{-}{r_D},\overset{-}{r_B},\overset{-}{r_E},\overset{+}{W}\right)+H_b^d(\overset{-}{r_D},\overset{-}{r_B})=H^s/P$$

$$(2)\quad D_h^d\left(\overset{+}{r_D},\overset{-}{r_B},\overset{-}{r_E},\overset{+}{W}\right)=D_b^s\left(\overset{-}{r_D},\overset{+}{r_B}\right)$$

$$(3)\quad B_h^d\left(\overset{-}{r_D},\overset{+}{r_B},\overset{-}{r_E},\overset{+}{W}\right)+B_b^d\left(\overset{-}{r_D},\overset{+}{r_B}\right)$$
$$=B_f^s\left(\overset{-}{r_B},\overset{+}{r_E},\overset{+}{K}\right)$$

$$(4)\quad E_h\left(\overset{-}{r_D},\overset{-}{r_B},\overset{+}{r_E},\overset{+}{W}\right)=E_f\left(\overset{+}{r_B},\overset{-}{r_E},\overset{+}{K}\right)$$

where r_i is the rate of return on asset i, and the signs above the arguments in the demand and supply functions refer to the signs of the partial derivatives. As a full-employment model, income is held constant at full utilization of resources. The general model has four endogenous variables, P, r_D, r_B, and r_E which are simultaneously determined by the four equations above and the definition of wealth.

This general system may be simplified considerably if two assumptions are made concerning the commercial banking sector. First, it will be assumed that the demand for high-powered money centers around the required ratio, denoted as ρ_0, and is affected inversely by the opportunity cost of reserves, i.e., r_B. Therefore H_b^d is replaced by the multiplicative expression $\rho(\rho_0, r_B)D_b^s$ in the market for high-powered money.[4] Sec-

[2] For exact signification of these conditions see Brainard and Tobin or, more recently, Santomero and R. D. Watson.

[3] The analysis below evaluates capital value at r_E equal to the marginal product of capital, following Tobin. (See fn. 7 below.) Accordingly P represents the price of current production of either commodities or capital in terms of money.

[4] The condition that ρ is a negative function of r_B can be interpreted as either the reserve function of an individual bank or an entire system. In the latter case the number of banks under high reserve requirements regulation, for example, Fed membership, is negatively related to the bond rate. See Section II, Part A, below.

ond, it will be assumed that the supply of deposits D_b^s is perfectly elastic at a given deposit rate r_D. This will be true if either 1) the supply of deposits exhibits constant returns to scale in high-powered money and bonds, so that the competitive deposit rate is $r_D=(1-\rho)r_b-c$, where c represents a constant per unit cost of producing deposits; or 2) the deposit rate r_D is fixed below the equilibrium value determined by (1) above.[5] The effect of these assumptions is to contract the four-equation system above to the following three:

$$(5)\quad H_h^d(r_D, r_B, r_E, W) + \rho(r_B, \rho_0)D_h^d(r_D, r_B, r_E, W) = H^s/P$$

$$(6)\quad B_h^d(r_D, r_B, r_E, W) + [1-\rho(r_B, \rho_0)] \times D_h^d(r_D, r_B, r_E, W) = B_f^s(r_B, r_E, K)$$

$$(7)\quad E_h^d(r_D, r_B, r_E, W) = K - B_f^s(r_B, r_E, K)$$

In this version of the model there are three unknowns, viz., r_B, r_E, and P, determined by the general equilibrium system of equations (5), (6), (7), and the definitional equation for wealth.[6] As the model is a comparative static full-employment framework, it appears appropriate to evaluate the system at a point where the financial market implies a full steady state in the economy. Tobin has shown that this occurs when the marginal product of capital in the real sector is equated to the cost of new equity. Accordingly, in Sections I and II the solution of the model is evaluated at the point where the rate of return on equity r_E is equal to the fixed marginal product.[7] This further reduces the analysis to two independent market equations, in two unknowns, viz., P and r_B. In Sections III and IV we explicity introduce uncertainty in r_E into the analysis.

It will prove convenient to consider equation (5), the high-powered money equation, and the sum of (5) and (6), which may be denoted the liquid asset equation. The latter contains the demand and supply of all assets in the system with the exception of equity, and has an equilibrium condition:

$$(8)\quad H_h^d(r_D, r_B, r_B, r_E, W) + D_h^d(r_D, r_B, r_E, W) + B_h^d(r_D, r_B, r_E, W) = B_F^s(r_B, r_E, K) + H^s/P$$

It can be shown (see Siegel) that as long as the banks' plus households' demand for bonds are positively related to the bond rate, the general equilibrium system is globally stable under any adjustment process of the endogenous variables to excess demand in the two markets. For example, excess demand for high-powered money could be viewed as inducing price level reduction while excess demand for bonds results in a reduction in the rate of interest paid to bond holders. The transformation of the system from equations (5) and (6) to the two-equation system (5) and (8) in no way changes the analytics of the model, but allows more intuition and ease of manipulation.

Converting equations (5) and (8) to excess demand notation is accomplished by letting H be the real excess demand for high-powered money, and A the real excess demand for liquid assets, as indicated in (8). The model may be written in reduced form:

$$(9)\qquad H\left(\overset{-}{r_B}, \overset{-}{r_E}, \overset{?}{r_D}, \overset{+}{\rho_0}, \overset{-}{H^s}, \overset{+}{P}\right)=0$$

$$(10)\qquad A\left(\overset{+}{r_B}, \overset{-}{r_E}, \overset{+}{r_D}, \overset{-}{H^s}, \overset{+}{P}\right)=0$$

The assumption of gross substitutes accounts for the deterministic signs on r_B and r_E. The ambiguity of $\partial H/\partial r_D$, denoted

[5] There is sufficient evidence in the literature to suggest that this is the case. Estimates of implicit yields and service returns on deposits by Robert Barro and Santomero, William Becker, and R. Startz all suggest that legal interest rate restrictions are somewhat effective.

[6] The evaluation of wealth excludes the monopoly rent accruing to the banking sector due to the restrictions on r_D and entry. On this subject, see Don Patinkin.

[7] This is consistent with evaluating the system at Tobin's "q" equal to one. As a full-employment model, this appears appropriate.

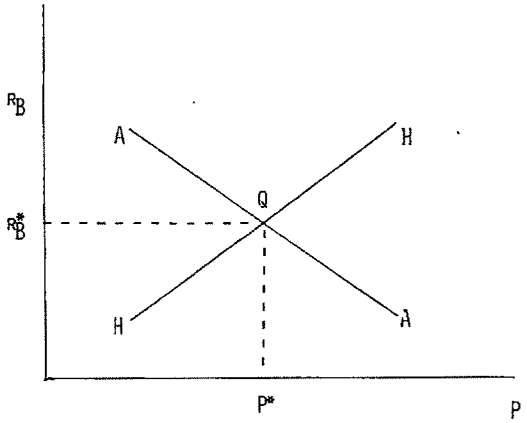


FIGURE 1. THE EQUILIBRIUM SOLUTION FOR THE DETERMINISTIC MODEL

henceforth as H_D, results from the offsetting effect of r_D increases on the households' and banks' demand for high-powered money. If high-powered money and deposits are net substitutes, the sign of H_D is negative. If they are complements, the opposite is implied. The signs of both H^s and P in equations (9) and (10) follow directly from the normality assumptions above, and both these equations are homogeneous of degree zero in H^s and P. This system of equations can be represented in two-dimensional space for a given value of r_E. As indicated in Figure 1, HH represents the locus of points where the market for high-powered money is in equilibrium and AA where the market for liquid assets is in equilibrium. Equilibrium in the system is obtained for a unique set of r_B^* and P^*.

B. *The Stochastic Behavior of the Model*

As noted at the outset, attention will center upon the stochastic characteristics of the model just presented. Two types of random disturbances can be considered. First, there may be a randomness associated with the market for high-powered money. Any change in the demand for high-powered money by commercial banks or currency by households causes a shift in this market. Supply variations by the central bank are also felt here. Shocks subsumed here are disturbances to the demand for money function, as Stephen Goldfeld has recently suggested or uncertainty on the supply side, as critics of the Federal Reserve often charge.[8] In Figure 1 this is indicated by shifts in the *HH* curve. Second, there may be randomness in the demand for liquid assets with an equal but opposite disturbance to the equity market. For example, shifts in the demand for equity at the expense of bonds are captured by this sort of disturbance. This is represented by shifts in the *AA* curve. Disturbances which affect both high-powered money and equity will shift both the *HH* and the *AA* curves. For example, a shift from currency to equity affects both markets. In general, the shocks to these two markets are correlated, with the degree of correlation and its sign depending upon the specific disturbance.

If the central bank has perfect knowledge of the excess demand functions (9) and (10), and the bond rate and price level without a lag, it can offset any shifts in the *HH* or *AA* curves by appropriate changes in high-powered money, the deposit rate, or the reserve ratio. It can be easily shown that the monetary authority has sufficient instruments to achieve any equilibrium. When the target variable, the price level, is only known with a lag, and the sources of the disturbances are not known, complete control is foregone. In this case, the central bank can only operate to minimize the fluctuations in the price level by structuring the financial environment so as to minimize the effect of such disturbances on the economy. For example, such policy changes may affect the *responsiveness* of the base money market to disturbances that alter interest rates. Hence the central bank can affect the slope of the *HH*. If, as shown in Section II below, the *AA* locus is unaffected by such a change in regulation, the effect can be seen with reference to Figures 2 and 3. Since the slope of *HH* equals $-H_p/H_B$, a policy change which makes the market for high-powered money more sensitive to the interest rate will flatten the *HH* curve to *H'H'*

[8] See, for example, Frost (1977), Albert Burger, Lionel Kalish, and Christopher Babb, and Bomhoff.

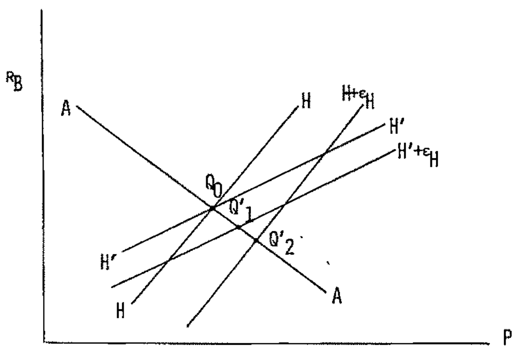


FIGURE 2. GRAPHICAL EXPOSITION OF DISTURBANCE IN THE HIGH-POWERED MONEY MARKET UNDER TWO FINANCIAL STRUCTURES

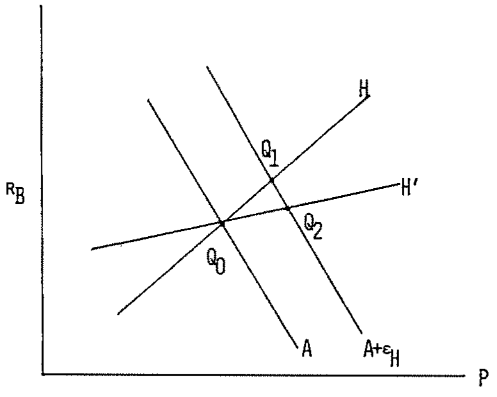


FIGURE 3. GRAPHICAL EXPOSITION OF DISTURBANCE OF THE DEMAND FOR LIQUID ASSETS UNDER TWO DIFFERENT FINANCIAL STRUCTURES

but leaves unchanged the horizontal displacement of the curves resulting from disturbances. A given disturbance to the market for high-powered money (ε_H), will thus result in a smaller increase in the price level (Q'_1) after the policy change than before (Q'_2). Hence such a policy shift is termed stabilizing with respect to disturbances in the market for high-powered money. On the contrary, as indicated in Figure 3, disturbances arising in the market for liquid assets (ε_A) will result in a greater change in the price level (Q_2) after the structural shift than before (Q_1). Hence such a policy is destabilizing with respect to disturbances in the market for liquid assets.

To consider the impact of stochastic disturbances analytically, the model must be specified more exactly. This is done in the text by linearizing the system around its equilibrium values of equations (9) and (10) as shown in Appendix A. In Appendix B, an alternative linearization is considered to illustrate the generality of the results. The basic linear model, in terms of deviations from equilibrium, may be written as

$$(11)\quad H_p\tilde{p}+H_B\tilde{r}_B=-H_D\tilde{r}_D-H_E\tilde{r}_E-H_\rho\tilde{\rho}_0 + H_p\,log\,\tilde{\bar{H}}+\varepsilon_H$$

$$(12)\quad A_p\tilde{p}+A_B\tilde{r}_B=-A_D\tilde{r}_D-A_E\tilde{r}_E+A_p\,log\,\tilde{\bar{H}}+\varepsilon_A$$

where p represents the *log* of the price level, the subscripted excess demands represent the partial derivatives evaluated at equilibrium, tildes represent deviations from the equilibrium values, and ε_H and ε_A are the random disturbances outlined above.[9] It will be assumed the expected values of both disturbances are zero, $E(\varepsilon_H)=E(\varepsilon_A)=0$, and $E(\varepsilon_H\varepsilon_A)=\sigma_{AH}$, $E(\varepsilon_H^2)=\sigma_H^2$, $E(\varepsilon_A^2)=\sigma_A^2$. Further, it will be assumed that the random disturbances terms are proportional to the inverse of the price level so that variations in the price level will not change the relative importance of the disturbances.[10]

For any given value of the disturbance terms ε_H and ε_A, the equilibrium price level, from equation (11) and (12) becomes

$$(13)\qquad \tilde{p}=\frac{(A_B\varepsilon_H-H_B\varepsilon_A)}{(H_pA_B-H_BA_p)}$$

[9] For example, if ε_C, ε_D, and ε_B represented demand shocks to the household's currency, deposits, and bonds, $\varepsilon_H=\varepsilon_C+\rho\varepsilon_D$ and $\varepsilon_A=\varepsilon_C+\varepsilon_D+\varepsilon_B$. The text deals with additive disturbances only. However, it should be noted that even if the disturbances were multiplicative an analogous condition would arise. The direction of the effect is identical for all shocks considered in the text.

[10] This assumption is employed so that the shocks are proportional to the size of the market. It is exact in the high-powered money market, and approximate for the liquid asset market.

and the variance of the *log* of the price level is

$$\text{(14)}\quad \text{var}(\tilde{p}) \equiv \sigma_p^2 = \frac{A_B^2\sigma_H^2 + H_B^2\sigma_A^2 - 2\sigma_{AH}A_B H_B}{(H_p A_B - H_B A_p)^2}$$

The focal point of the analysis will be the effect of the proposed regulatory changes on the variance term of equation (14). Given the full-employment nature of the model, the concern with the variance in price follows directly. If the effect of regulatory changes is an increase in the variance of p, then structural change reduces effective control of the economy by the monetary authority, and in this sense is a detrimental shift. Of course a decrease in the variance of p is salutary.[11] It remains to outline these proposed shifts and their effect on the variance term.

II. The Effect of Proposed Policy Changes

The method of analysis in this study is to determine how several regulatory changes affect the slope of the equilibrium curves and hence the variance of the price level. Three potential changes in the structural environment of the banking system are possible. As indicated previously, all are provided for in the recent Depository Institutions Deregulation and Monetary Control Act of 1980 approved both by the Congress and the President as Public Law 96-221. These are

1) a change in the required reserve ratio of member and nonmember banks,

2) interest payment on reserves held by financial institutions, and

3) interest payment on demand deposits.

The first two were enacted as a part of a comprehensive program to address the rapidly declining membership in the Federal Reserve while the third was in response to competitive pressures from savings banks and selective commercial banks issuing interest-bearing checking accounts. This study will analyze how such legislation and future variations in these regulations affect macroeconomic control and how they may be used in combination to improve macro stability.

The procedure used to evaluate the effect of the three policy changes will be the same. In each case the total differential of equation (14) with respect to the proposed policy change is taken. The result will be a change of the slope of the *HH* curve, whose graphical interpretation was presented in the last section.

A. *Changes in Reserve Ratios*

The first change to be considered is the effect of a change in required reserve ratios. Above it was assumed that the reserve function was dependent upon the rate of return on bonds and the required ratio. If it is further assumed that a change in required reserves, *ceteris paribus*, increases actual reserves by an equal amount, at least on the margin, then the reserve function may be written as

$$\text{(15)}\qquad \rho = \rho(\rho_0, r_B) = \rho_0 + \rho(r_B)$$

The effect of a small change in the required reserve ratio ρ_0 on the variance of the *log* of the price level is formally equal to $\partial\sigma_p^2/\partial\rho_0$. In order to evaluate this expression some simplification is necessary. If the demand functions are assumed linear in rates of returns and wealth, the expression becomes (see Appendix A. II. for derivation)

$$\text{(16)}\quad \frac{\partial\sigma_p^2}{\partial\rho_0} = 2\left[A_B\frac{\partial H_B}{\partial\rho_0}\right]\Big[\sigma_H^2 A_B A_p + \sigma_A^2 H_B H_p - \sigma_{AH}(A_B H_p + H_B A_p)\Big] \div (H_p A_B - H_B A_p)^3 - \frac{2\sigma_p^2 A_B\dfrac{\partial H_p}{\partial\rho_0}}{(H_p A_B - H_B A_p)}$$

[11] An alternative way to analyze the regulatory change is to endogenize r_E rather than p. The results are qualitatively similar, however.

where

$$\text{(17a)}\quad \frac{\partial H_p}{\partial \rho_0} = -\frac{\partial \overset{+}{D}{}_h^d}{\partial W}\left[1+\frac{\partial \rho}{\partial r_B}\frac{dr_B}{d\rho_0}\right]<0$$

$$\text{(17b)}\quad \frac{\partial H_B}{\partial \rho_0} = \frac{\partial \overset{-}{D}{}_h^d}{\partial r_B}$$

$$+\overset{-}{\frac{\partial \rho}{\partial r_B}}\left[2\overset{-}{\frac{\partial D_H^d}{\partial r_B}}\frac{d\overset{+}{r}\;{}_B}{d\rho_0}+\frac{\partial \overset{+}{D}{}_h^d}{\partial W}\frac{d\overset{+}{W}}{d\rho_0}\right]\overset{?}{<}0$$

Equation (17a) captures the change in the sensitivity of the excess demand for high-powered money to the price level due to the wealth effect caused by the drop in prices and the increase in real high-powered money. A decrease in H_p means that the demand for high-powered money is becoming more sensitive to price changes. Hence any shock to the system requires a greater change in prices to bring demand and supply back to equilibrium and the variance of prices must increase. Graphically, this is equivalent to a larger horizontal shift of the equilibrium curves for any given real shock. Therefore the effect of the second term in equation (16) is to increase the variance of prices.

The first term of equation (17b) is negative, whereas the sign of the second term is ambiguous since the bond rate rises and real wealth increases in response to an increase in ρ_0. The first term is the direct effect of an increase in ρ_0 on the sensitivity of high-powered money to the bond rate. An increase in ρ_0 causes a rise in the real amount of reserves. For a given interest sensitivity of deposits, the interest sensitivity of total reserves increases, and hence the direct effect of a rise in ρ_0 is negative. The second term captures the indirect effects of a shift in ρ on the bond rate and overall wealth. The first of these is an offsetting effect, as the bond rate will rise with ρ. On the other hand, real wealth increases as prices decline with an increase in ρ. If these indirect effects do not predominate, then $\partial H_B/\partial\rho_0$ is negative. This is indicative by a flatter HH curve as depicted in Figure 2 and 3. The AA curve is unaffected since the reserve requirement under our linear demand assumption does not affect either A_B or A_p. The sign of equation (16), then, is opposite in direction to the sign of the second bracketed term in the numerator, which may now be analyzed.

Consider the case of a disturbance in each market alone. If $\sigma_A^2=0$, the expression reduces to

$$\text{(18)}\quad \frac{\partial \sigma_p^2}{\partial \rho_0} = \frac{2A_B(\partial H_B/\partial \rho_0)A_BA_p\sigma_H^2}{(H_pA_B-H_BA_p)^3}$$

$$-\frac{2\sigma_p^2A_B(\partial H_p/\partial\rho_0)}{(H_pA_B-H_BA_p)}\begin{matrix}>\\[-4pt]-\\[-4pt]<\end{matrix}0$$

An increase in the reserve ratio has an ambiguous effect on the variance of prices since the increased interest sensitivity of high-powered money is offset by the decreased sensitivity of the excess demand for high-powered money to prices. Figure 2 demonstrates the dampening effect on price variability caused by the change in H_B. Appendix B derives the exact conditions for determining the sign of (18) under an alternative linearization. In the case where only the liquid asset market is subject to disturbances, i.e., where $\sigma_H^2=0$, equation (16) becomes

$$\text{(19)}\quad \frac{\partial \sigma_p^2}{\partial \rho_0} = \frac{2A_B(\partial H_B/\partial \rho_0)\sigma_H^2H_pH_B}{(H_pA_B-H_BA_p)^3}$$

$$-\frac{2\sigma_p^2A_p(\partial H_p/\partial\rho_0)}{(H_pA_B-H_BA_p)}>0$$

Here, the price sensitivity and interest shifts are reinforcing and the variance of prices unambiguously rises. Figure 3 shows the effect on price variability of an increase in the interest sensitivity of high-powered money. Finally, in the general case when both variances are nonzero and the covariance terms of equation (16) become relevant, the effect of a shift in ρ is ambiguous. The larger the σ_H^2 term the more likely a positive shift in ρ will reduce overall variance in prices.[12]

Generalizing from above, one may now summarize the result of a (positive) shift in ρ_0, as captured in equation (16). If disturbances prevail primarily in the market for high-powered money, for instance, resulting from demand shifts among liquid assets holding their total constant, the effect of a rise in a reserve requirement is ambiguous on the variance of the price level. On the other hand, if the disturbances are primarily in the demand for total liquid assets (i.e., shifts between equity and bonds), the effect of a rise in ρ causes greater price variability for a given disturbance in this market.

B. *Interest Payments on Required Reserves*

The proposal for the payment of interest on all reserves was not incorporated into the 1980 legislation. However the Act does permit the Federal Reserve to pay interest on marginal, or supplemental, reserves.[13] As the issue of interest payment on reserves is still subject to much debate and is likely to play a role in future legislative or administrative changes, the case of a straightforward payment on reserves will be treated here. Such a variation in regulation can be analyzed easily in the framework developed above. The payment of interest should be expected to increase the average reserve ratio on demand deposits. The effect of interest payments would be the greater reluctance of existing member banks to have a negative free reserve position, with its system penalties. If payment of interest is made on all reserves, there would also be a smaller incentive for banks to avoid excess reserves, as well as greater incentives to avoid a deficiency.

The effect of increasing interest returns to reserves, denoted r_H, may be written formally as

$$\text{(20)} \qquad \frac{\partial\sigma_p^2}{\partial r_H} = \frac{\partial\sigma_p^2}{\partial\rho}\,\frac{d\rho}{dr_H}$$

with the first term equal to equation (16) above, and $d\rho/dr_H > 0$. Hence, the institution of interest payment on reserves has an uncertain effect on the variance of the price level if disturbances occur primarily among the markets for currency, bonds, and deposits, but increases the variance if disturbances are primarily between the market for bonds and equity.

C. *Interest Payments on Demand Deposits*

The legislation allows interest payments at a government-administered ceiling rate[14] on demand deposits. Under Title II of the Act, a Depository Institutions Deregulation Committee is established to set rules for the eventual phase out of interest rate ceilings. The effect of the payments on demand deposits impacts directly upon the deposit market and the household, whereas the other regulatory changes affected the financial institutions first. However, the method of analysis proceeds in a similar fashion.

The effect of these changes on the variance of prices is given by

$$\text{(21)} \qquad \frac{\partial\sigma_p^2}{\partial r_D} = \left[2A_B\frac{\partial H_B}{\partial r_D}\right]\Big[\sigma_H^2 A_B A_p$$

[12] The effect of the covariance between the shocks depends upon the sign of $(A_B H_p + H_B A_p)$. If the absolute value of the slope of HH is greater than that of the AA locus, then $(A_B H_p + H_B A_p)$ is positive. The effect of a positive covariance between the disturbances, in this case, increases the anticipated variance of the price level associated with the structural change under consideration.

[13] "The Board is also given the authority, upon the affirmative vote of not less than five members, to impose a supplemental reserve requirement on every depository institution of not more than 4% of its total transactions accounts... . That account shall receive earnings to be paid by the Federal Reserve Banks during each calendar quarter at a rate not more than the rate earned on the securities portfolio of the Federal Reserve System on the previous calendar quarter." [Title I Public Law 96-221.]

[14] For the present framework it is required that the increase in r_D is insufficient to free the market for demand deposits from the government ceiling regulation. The model in the text assumes that the r_D constraint is binding in the deposit market throughout the experiment. If, however, the market became free from constraint as r_D increased, the sign but not the magnitude of the effect would still be the same.

$$+\sigma_H^2 H_p H_B - \sigma_{AH}(A_B H_p + H_B A_p)\big]$$
$$\div (H_p A_B - H_B A_p)^3$$
$$-\frac{2\sigma_p^2 A_B \dfrac{\partial H_p}{\partial r_D}}{(H_p A_B - H_B A_p)}$$

where

$$\text{(22a)}\quad \frac{\partial H_p}{\partial r_D} = -\frac{\partial \overset{+}{D_h^d}}{\partial W}\,\frac{\partial \overset{-}{\rho}}{\partial r_B}\,\frac{d\overset{-}{r_B}}{dr_D} < 0$$

$$\text{(22b)}\quad \frac{\partial H_B}{\partial r_D} = \frac{\partial \overset{-}{\rho}}{\partial r_B}\left[\frac{\partial \overset{+}{D_h^d}}{\partial r_D} + 2\frac{\partial \overset{-}{D_h^d}}{\partial r_B}\,\frac{d\overset{-}{r_B}}{dr_D}\right.$$
$$\left. + \frac{\partial \overset{+}{D_h^d}}{\partial W}\,\frac{d\overset{+}{W}}{dr_D}\right] < 0$$

Unless deposits and high-powered money are extreme complements (see Appendix A.III.), the bond rate will fall in response to a rise in deposit rates and hence H_p will fall. Therefore, for reasons identical to those discussed in the case of a change in reserve requirements, this reduction in H_p caused by the wealth effect of falling prices, will increase the variability of prices.

The first term in equation (22b) represents the sensitivity of the average reserve ratio to the bond rate times the shift in deposits due to a change in the deposit rate. This is unambiguously negative. The second and third terms capture the indirect effects, as was the case in equation (17) above. The second term depends upon the sign of dr_B/dr_D, while the third is determined by dW/dr_D. Ruling out extreme complementarity and extreme substitutability between high-powered money and deposits,[15] both the bond rate and the price level will fall in response to a rise in the deposit rate. These terms are of identical sign to the primary effect and reinforce the direct effect of r_D. Therefore, equation (22b) will be negative and the *HH* curve become flatter when deposits rates rise. Under the linearity assumptions, the slope of *AA* is again unaffected by a rise in r_D. Hence, the effect of r_D on the variance of prices is, as before, dependent upon the nature of the disturbances facing the financial structure.

Therefore, the analysis of the change in the deposit rate is qualitatively identical to that of a change in the reserve requirement. A rise in deposit rates will have an ambiguous effect on the variance of the price level for disturbances in the high-powered money market but will lead to greater price fluctuations if the disturbances center in the market for liquid assets.

III. Financial Interaction with the Real Sector

Thus far the analysis has examined the effect of regulatory reforms on the system with financial disturbances only. This section expands the focus of the analysis by considering the impact of financial reforms on the system's behavior when faced with real shocks. To analyze disturbances in the real sector, the financial sector model is converted to (r_E, P) space. Equations (11) and (12) can be solved for p in terms of r_E and the disturbance terms are

$$\text{(23)}\qquad \tilde{p} = F_E r_E + \tilde{\varepsilon}_F$$

where $$F_E = \frac{H_B A_E - A_B H_E}{H_p A_B - H_B A_p} > 0$$

and $$\tilde{\varepsilon}_F = \frac{A_B \tilde{\varepsilon}_H - H_B \tilde{\varepsilon}_A}{H_p A_B - H_B A_p}$$

Equation (23) can be interpreted as the *mutatis mutandis* financial locus and is depicted by the upward sloping locus *FF* in Figure 4. The slope coefficient is denoted F_E, and the effect of the disturbances is captured in the ε_F term. Note that the bond rate changes *mutatis mutandis* so as to maintain equilibrium in the financial market along the *FF* locus.

[15] See the Appendix A for the exact condition.

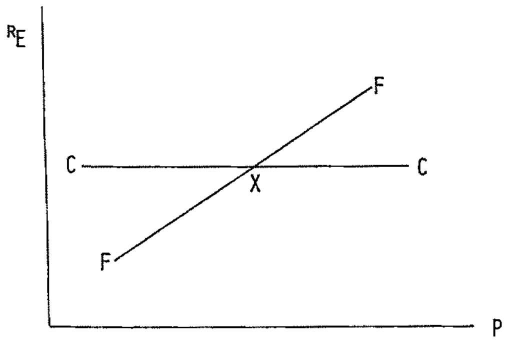


FIGURE 4. GENERAL EQUILIBRIUM SOLUTION OF THE MODEL

To the financial sector we add a commodity market locus. As interest centers around the long-run effects of regulatory change, commodity demand functions are assumed to derive from some intertemporal utility maximization procedure. These models[16] often result in a flow commodity or output demand function C^d, which is a function of only the equity rate.[17] Output supply C^s is exogenously determined at its equilibrium level. Both the supply and demand for output are subject to zero-mean real shocks, ε_R^d and ε_R^s. The equilibrium condition for the output market can be written as

(24)

$$C^s = C^d(r_E) + \tilde{\varepsilon}_R, C_E = \frac{\partial C^d}{\partial r_E} < 0, \tilde{\varepsilon}_R = \tilde{\varepsilon}_R^d - \tilde{\varepsilon}_R^s$$

The locus of points where the commodity market is in equilibrium is represented by the horizontal line CC in Figure 4.

Linearizing equation (24) above and substituting into equation (23) yields the following solution for the price level disturbances and variance

$$(25)\quad \tilde{p} = -\frac{F_E}{C_E}\tilde{\varepsilon}_R + \tilde{\varepsilon}_F$$

$$(26)\quad \sigma_p^2 = (F_E/C_E)^2\sigma_R^2 - 2(F_E/C_E)\sigma_{RF} + \sigma_F^2$$

Abstracting from the financial disturbances that have been treated above, the effect of regulatory change on the variance of price due to real shocks can be written as

$$(27)\qquad \frac{\partial \sigma_p^2}{\partial x_0} = \frac{2F_E}{C_E^2} \cdot \frac{\partial F_E}{\partial x_0}\sigma_R^2$$

where x_0 is any regulatory change. The sign of equation (27) is therefore determined by the sign of $\partial F_E/\partial x_0$, the effect that the regulatory change has on the slope of the *FF* locus. It is obvious from inspection, and the work of Poole in an analogous context, that the disturbances originating in the real sector have a greater impact on the price level when the *FF* locus in Figure 4 is shallower.

The procedure used to evaluate the effect of the regulatory change on the variance of the price level due to real disturbances will be the same as that conducted in Section II. The total differential of equation (26) will be analogous to equation (16) for financial disturbances. Accordingly, the result of a change in ρ_0 on the slope of F_E, which determines the variance from real disturbances can be derived as

(28)

$$\begin{aligned} \frac{\partial F_E}{\partial \rho_0} = A_B\big[&(\partial H_B/\partial \rho_0)(H_p A_E - A_p H_E) \\ &-((\partial H_E/\partial \rho_0) + F_E(\partial H_p/\partial \rho_0)) \\ &\times (H_p A_B - H_B A_p)\big] \\ &+ (H_p A_B - H_B A_p)^2 > 0 \end{aligned}$$

where all terms are defined as above.

[16] See Rudiger Dornbusch and Jacob Frenkel for a discussion of the long-run commodity demand function.

[17] The model presented in the text abstracts from the price effects on commodity demand. As indicated, this is the result of our interest in long-run effects, and the irrelevency or ambiguity of the price effect in the long run. If the model had considered the price effect in the commodity locus, all qualitative results of the model would remain but a substantial increase in complexity would result. The proofs for this version of the model are available from the authors.

The sign of equation (28) is unambiguously positive. Hence the slope of the *FF* curve rises in the case of an increase in ρ_0 and, for a given disturbance in the real sector, the variance of the price level will increase. As noted above in Section II, Part B, an increase in the interest paid on reserves will have identical qualitative effects to an increase in reserve requirements.

An increase in the deposit rate, the third structural change, can be analyzed in a similar fashion to ρ_0 above. Taking the derivative of the F_E locus with respect to r_D under the assumptions that obtained above results in

(29)

$$\frac{\partial F_E}{\partial r_D}=\frac{A_B(H_pA_E-H_EA_p)(\partial H_B/\partial r_D)}{(H_pA_B-H_BA_p)^2}>0$$

As in the case of reserve requirements, a rise in the deposit rate unambiguously increases the variance of the price level for any given disturbance in the real sector.

IV. Integration of Real and Monetary Disturbances

Section II above indicated the results of regulatory change on the overall variance of prices associated with stochastic disturbances in the monetary sector. It indicated that if wealth effects are ignored, the effect on the variance of the price level of shocks to base money are decreased by positive changes in ρ_0 and r_D. Also, because payment of interest on reserves increases the reserve ratio, an increase in this rate, r_H, also decreases overall variance. The effect of shocks between bonds and equity is definitive, but opposite in sign to that of the base money disturbances. It is straightforward to analyze the set of both reserve requirement changes and deposit rate changes that could be combined so that the variance of the price level remains constant for any set of disturbances in the financial market. The slope of this isovariance curve, which shall be termed *MM* is

$$(30)\qquad \left.\frac{\delta\rho_0}{\delta r_D}\right|_{MM}=-\frac{\partial\sigma_p^2}{\partial r_D}\bigg/\frac{\partial\sigma_p^2}{\partial\rho_0}$$

Substituting equations (16) and (21) into equation (30) and ignoring indirect effects of changes in r_B and induced wealth effects yields the simple form

$$(30')\qquad \left.\frac{\delta\rho_0}{\delta r_D}\right|_{MM}=-\frac{\partial\rho}{\partial r_B}\frac{\partial D_h^d}{\partial r_D}\bigg/\frac{\partial D_h^d}{\partial r_B}$$

Further, if it is assumed that deposit demand may be regarded as primarily a function of the difference between r_B and r_D, so that $\partial D_h^d/\partial r_D \cong -\partial D_h^d/\partial r_B$, then equation (30) reduces to

$$(30'')\qquad \left.\frac{\delta\rho_0}{\delta r_D}\right|_{MM}=\frac{\partial\rho}{\partial r_B}$$

This indicates that if the monetary authority were to change the average reserve ratio and the deposit rate in the same ratio as the rate of responsiveness of the marginal aggregate reserve ratio to change in the market interest rate, then fluctuations in the price level in the economy would neither increase nor decrease due to disturbances in the monetary sector.

Just as in the case of the financial market, Section III implies that a locus of isovariance for real disturbances can be traced for different values of the policy instruments ρ_0 and r_D. The slope of this locus, denoted *RR*, can be formed by substituting (28) and (29) into (31) below:

$$(31)\qquad \left.\frac{\delta\rho_0}{\delta r_D}\right|_{RR.}=-\frac{\partial\sigma_p^2}{\partial r_D}\bigg/\frac{\partial\sigma_p^2}{\partial\rho_0}$$

$$=\frac{\partial\rho}{\partial r_B}\frac{\partial D_h^d}{\partial r_D}\bigg/\left(\frac{\partial D}{\partial r_B}-q\right)<0$$

where

$$q=\frac{(\partial D_h^d/\partial r_E)(H_pA_B-H_BA_p)}{(H_pA_E-H_EA_p)}>0$$

As displayed in Figure 5, it can be seen that the isovariance locus for disturbances in the commodity market is shallower than the *MM* locus in (ρ_0, r_D) space.

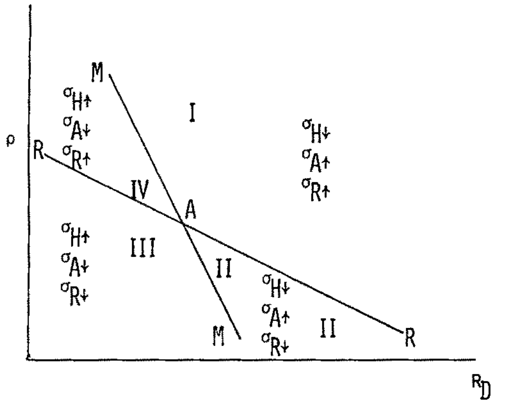


FIGURE 5. THE EFFECT OF REGULATORY CHANGE ON OVERALL PRICE VARIANCE

σ_H = VARIANCE DUE TO MONETARY DISTURBANCES

σ_A = VARIANCE DUE TO LIQUID ASSET DISTURBANCES

σ_R = VARIANCE DUE TO REAL OR COMMODITY MARKET DISTURBANCES

As indicated, the (ρ_0, r_D) plane is divided into four regions. Starting from the existing regulatory structure of point A, the monetary authority can move in any direction. If it believes that stochastic shocks arise primarily in the high-powered money market, the effect of these disturbances on the price level can be mitigated, compared to point A, by moving into regions I or II. However, the effect of disturbances in the real sector are increased in region I, and decreased in II and III only. Therefore, the movement of the instruments ρ_0 and r_D to the second region decreases both real and monetary disturbances. If, on the other hand, financial disturbances arise primarily from liquid asset shifts, then the direction of movement of σ_A would favor region III over region II.

It should be noted that one cannot say that all points in a particular region dominate all points in another. For instance, if monetary disturbances overwhelm all other sources of variance, the true optimum may be in region I where the effect of monetary disturbances may be greatly mitigated at the cost of a slight worsening of the effect of real disturbances. However, region II does represent a conservative policy posture area if the Federal Reserve perceives both base money and real disturbances to be important but has limited knowledge as to the relative weights.

The foregoing analysis also reflects upon the debate over the appropriate way to lessen the cost of reserves to the banking industry. Two methods are provided for in the existing legislation. One would reduce the reserve ratio directly, while the second would indirectly increase it by paying interest on reserves. The first of these alternatives would exacerbate the effect of high-powered money disturbances and reduce the effects of real shock since it would move the economy into region III. On the other hand, the second alternative would have exactly the opposite effect, by moving the system into region I. If the reduction in the reserve requirement is coupled with increases in the interest on demand deposits, the alternative which reduces ρ_0 has the potential of moving into region II, while an increase in reserve requirements keeps the system in I. Therefore, if one views the real sector as a major source of disturbances, the alternative of reducing reserve levels and raising the rate of interest on deposits dominates an increase in reserves, for the purpose of economic stabilization.

V. Summary and Conclusions

This paper analyzed the impact of various regulatory changes in the banking area on the stochastic behavior of prices in a macroeconomic framework. It indicated that, on the financial side, holding the mean of the price level constant, regulations such as an increase in required reserves or interest on demand deposits reduce the impact of base money disturbances on the variability of the price level, while increasing the response of the price level to liquid asset shocks.

The study also investigated the impact of such regulation on the system's response of real or commodity market shifts. Here it was found that increases in required reserves and interest paid on demand deposits generally increased the fluctuations of the price level caused by shocks to the real sector. It was concluded that, unless random shifts between equity and other financial assets predominate, selective lowering of reserve requirements and raising of deposit rates would be beneficial for economic stabilization.

Appendix A—The Basic Equations of the Model

I. The basic structure of the model indicated in equations (9) and (10) is

$$\text{(A1)} \quad \frac{H}{P}\begin{bmatrix} H_B & H_p \\ A_B & A_p \end{bmatrix}\begin{bmatrix} r_B \\ \log P \end{bmatrix}=0$$

where

$$H_B \equiv \frac{\partial H_h^d}{\partial r_B}+\frac{\partial \rho}{\partial r_B}D_h^d+\rho\frac{\partial D_h^d}{\partial r_B}$$

$$H_p \equiv -\left[\frac{\partial H_h^d}{\partial W}+\rho\frac{\partial D_h^d}{\partial W}\right]\frac{H}{P}+\frac{H}{P}$$

$$A_B \equiv \frac{\partial H_h^d}{\partial r_B}+\frac{\partial D_h^d}{\partial r_B}+\frac{\partial B_h^d}{\partial r_B}-\frac{\partial B_f^S}{\partial r_B}$$

$$A_p \equiv -\left[\frac{\partial H_h^d}{\partial W}+\frac{\partial D_h^d}{\partial W}+\frac{\partial B_h^d}{\partial W}\right]\frac{H}{P}+\frac{H}{P}$$

II. The general form of the effect of exogenous structural shifts on the variance of *log p* is

$$\text{(A2)} \quad \frac{\partial \operatorname{var}(\tilde{p})}{\partial x_0}=2\left[A_B\frac{\partial H_B}{\partial x_0}-H_B\frac{\partial A_B}{\partial x_0}\right]$$

$$\times\left[\sigma_H^2 A_B A_p+\sigma_A^2 H_p H_B-\sigma_{AH}\right.$$

$$\left.\times(A_B H_p+H_B A_p)\right]$$

$$\div(H_p A_B-H_B A_p)^3$$

$$-2\operatorname{var}(\tilde{p})\left[A_B\frac{\partial H_p}{\partial x_0}-H_B\frac{\partial A_p}{\partial x_0}\right]$$

$$\div(H_p A_B-H_B A_p)$$

Given the relationship of the error to the price level (see fn. 11), H/P does not appear in (A2). Equations (16) and (20) are special cases of this general form.

III. The result of an increase in r_D on r_B and p from I above is

$$\text{(A3)} \quad \frac{dr_B}{dr_D}=\frac{H_D A_p-H_p A_D}{H_p A_B-A_p H_B}<0$$

and
$$\frac{dp}{dr_D}=\frac{A_D H_B-A_B H_D}{H_p A_B-H_B A_p}<0$$

Unless, high-powered money and deposits are extreme complements, i.e, $H_D \gg 0$, the bond rate must fall in response to a rise in the deposit rate. Similarly, if H and D are not extreme substitutes, i.e., $H_D \ll 0$, the price level must also fall.

Appendix B—A Simplified Model with Linear Homogenity in Wealth

This section derives the conditions for determining the sign of (16) under an alternative linearization where household asset demands are proportional to wealth, so that

$$\text{(A4)} \quad C_h^d=a_c(r_d,r_b,r_e)W$$

$$D_h^d=a_d(r_d,r_b,r_e)W$$

$$B_h^d=a_b(r_d,r_b,r_e)W$$

$$E_h^d=a_e(r_d,r_b,r_e)W$$

where $W=\bar{K}+H/P$, and $a_c+a_d+a_b+a_e=1$. For simplicity, assume full equity financing of the existing capital stock and an exogenous reserve ratio ρ. The first assumption implies that bonds held by banks are entirely loans to households, so that a_d is negative and

$$\text{(A5)} \quad a_c+a_d+a_b=a_c+\rho a_d$$

It is immediate that

$$\text{(A6)} \quad A_B=(a_c'+a_d'+a_b')W$$

$$H_B=(a_c'+\rho a_d')W$$

$$A_p=H_p=(H/P)a_e$$

where $\quad a_i'=\partial a_i/\partial r_B$

Then (14) can be written

$$\text{(A7)} \qquad \sigma_p^2 = \frac{\hat{\sigma}_A^2 Z^2 + \hat{\sigma}_H^2 - 2Z\hat{\sigma}_{AH}}{[(H/P)a_e(Z-1)]^2}$$

where $Z = H_B/A_B$

If we assume (see fn. 10),

$$\text{(A8)} \quad \varepsilon_H = (H/P)\hat{\varepsilon}_H \text{ and } \varepsilon_A = (H/P)\hat{\varepsilon}_A$$

then

$$\text{(A9)} \qquad \sigma_p^2 = \frac{\sigma_p^2 Z^2 + \sigma_H^2 - 2Z\sigma_{AH}}{[a_e(Z-1)]^2}$$

If we examine the case of shocks to the excess demand for high-powered money only, i.e., $\sigma_A^2 = 0$, then

(A10)

$$\frac{\partial \sigma_p^2}{\partial \rho} = \frac{-2\sigma_H^2}{[a_e(Z-1)]^3}\left[a_e \frac{dZ}{d\rho} + (Z-1)a_e' \frac{dr_b}{d\rho}\right]$$

where

$$\text{(A11)} \quad \frac{dZ}{d\rho} = \frac{a_d'}{a_c' + a_d' + a_b'} < 0$$

$$\text{(A12)} \quad \frac{dr_b}{d\rho} = \frac{a_d}{[(1-\rho)a_d' + a_b']} > 0$$

The first term in the bracket of (A10) is negative, reflecting the increased interest sensitivity of the demand for high-powered money. The second term measures the shift in the equity fraction resulting from the increase in the reserve ratio. The equity fraction drops since prices fall and wealth in the form of high-powered money increases.

When (A11) and (A12) are substituted into (A10), it is immediate that

$$\text{(A13)} \quad \frac{\partial \sigma_p^2}{\partial \rho} \gtreqless 0 \text{ iff } \frac{a_e a_d'}{a_c' + a_d' + a_b'} + a_d \gtreqless 0$$

This condition reduces to

$$\text{(A14)} \quad \frac{\partial \sigma_p^2}{\partial \rho} \gtreqless 0 \text{ iff } a_e'/a_e \gtreqless a_d'/a_d$$

Hence an increase in the reserve ratio will *increase* the variability of prices due to disturbances in the demand or supply of high-powered money if and only if the (absolute value of) the elasticity of deposit demand is less than the elasticity of equity demand with respect to the bond rate.

In the case where $\sigma_H^2 = 0$,

$$\text{(A15)} \quad \frac{\partial \sigma_p^2}{\partial \rho} = \frac{2Z\sigma_A^2}{[a_e(Z-1)]^3} \cdot \frac{-a_e a_d' - a_d(a_c' + \rho a_d')}{a_c' + a_d' + a_b'} > 0$$

which is unambiguously positive, as is also demonstrated in the text. Therefore, an increase in the reserve ratio always increases the variability of prices if the shocks are due to shifts in the demand for liquid assets.

This Appendix demonstrates that under an alternative linearization where asset demands are proportional to wealth, the identical qualitative results found in the text apply. In addition, the exact condition for determining the sign of $\partial\sigma_p^2/\partial\rho$ due to high-powered money shocks is derived.

REFERENCES

R. J. Barro and A. M. Santomero, "Household Money Holdings and the Demand Deposit Rate," *J. Money, Credit, Banking,* May 1972, *25*, 397–413.

W. E. Becker, Jr., "Determinants of the United States Currency-Demand Deposit Ratio," *J. Finance,* Mar. 1975, *30*, 57–74.

F. J. Bomhoff, "Predicting the Money Multiplier," *J. Monet. Econ.,* Oct. 1977, *3*, 325–46.

W. C. Brainard, "Financial Intermediaries and a Theory of Monetary Control," *Yale Econ. Essays,* Nov. 2, 1964, *4*, 431–82.

A. E. Burger, L. Kalish II, and C. T. Babb, "Money Stock Control and its Implication for Monetary Policy," *Fed. Reserve Bank St. Louis Rev.,* Oct. 1971, *53*, 6–22.

R. G. Davis, "Short Run Targets for Open Markets Operations," in Board of Governors of the Federal Reserve System, *Open Market Policies and Operating Procedures-Staff Studies,* 1971, 37–69.

R. Dornbusch and J. Frenkel, "Inflation and Growth: Alternative Approaches," *J. Money, Credit, Banking,* Feb. 1973, *5,* 141-156.

P. Frost, "Short Run Fluctuations in the Money Multiplier and Monetary Control," *J. Money, Credit, Banking,* Feb. 1977, *9,* 165-81.

R. A. Gilbert, "Effects of Interest on Demand Deposit: Implications of Compensating Balances," *Fed. Reserve Bank St. Louis Rev.,* Nov. 1977, *59,* 8-15.

S. M. Goldfeld, "The Case of Missing Money," *Brookings Papers,* Washington 1976, *3,* 683-789.

R. Holbrook and H. Shapiro, "The Choice of Optimal Intermediate Economic Targets," *Amer. Econ. Rev. Proc.,* May 1970, *60,* 40-46.

I. Kaminow, "Why Not Pay Interest on Member Bank Business Reserves?," *Fed. Reserve Bank Philadelphia Rev.,* Jan. 1975, 3-9.

B. Klein, "Competitive Interest Payments on Bank Deposits and the Long Run Demand for Money," *Amer. Econ. Rev.,* Dec. 1974, *64,* 931-49.

D. Patinkin, "Money and Wealth, A Review Article," *J. Econ. Lit.,* Dec. 1969, *7,* 1140-60.

W. Poole, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quart. J. Econ.,* May 1970, *84,* 197-216.

A. M. Santomero and R. D. Watson, "Determining an Optimal Capital Standard for the Banking Industry," *J. Finance,* Sept. 1977, *32,* 1267-82.

J. J. Siegel, "A General Macro-Equilibrium Approach to Deposit Creation," unpublished paper, Univ. Pennsylvania, 1977.

R. Startz, "Implicit Interest on Demand Deposit," *J. Monet. Econ.,* Oct. 1979, *5,* 515-34.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money Credit, Banking,* Feb. 1969, *1,* 15-29.

______**and W. C. Brainard,** "Financial Intermediaries and the Effectiveness of Monetary Controls," *Amer. Econ. Rev. Proc.,* May 1963, *53,* 383-400.

Federal Reserve Board of Governors, "The Impact of Interest on Demand Deposits," Staff Study, Washington, Jan. 31, 1977.

Federal Reserve Bulletin, "The Depository Institutions Deregulation and Monetary Control Act of 1980," Washington, June 1980, *66,* 444-53.

The Economics of Risks to Life

By W. B. ARTHUR*

One of the more difficult questions the economist faces is how to assess activities—engineering projects, safety procedures, medical advances—that raise or lower risks to human life. It is clear that in most situations proper safety should be a matter of degree: engineering constructions should neither be infinitely solid nor built too close to their limits of strength. But how much safety should we strive for? What meaning can be given to phrases such as "the value of life" or "the cost of hazards to life?" And what are the economic consequences of the fact that mortality risks are gradually falling—that life is lengthening?

Two methods for evaluating mortality risks are currently available. The *human capital* approach assesses increased risk by earnings forgone through incapacitation or premature death.[1] It has the appealing property that it is *actuarial*: it uses full age-specific accounting to evaluate changes in mortality. But in spite of the precise dollar figures it offers, it is founded at best on thin logic. By concentrating purely on wage or *GNP* loss, it ignores, for example, the individual's own desire to live. Under human capital a medical breakthrough that prolonged life from seventy to eighty years would have no particular social justification—it would not raise *GNP*.

The *willingness-to-pay* method does recognize the natural desire to live longer. Under this approach, a scheme that increased life from seventy to eighty years would be socially justified if those who benefited were willing, in theory at least, to pay more for their extra years than the cost of the scheme. This literature *is* grounded solidly on welfare theory logic—it proceeds deductively from commonly accepted assumptions.[2] But it is not actuarial: it would have difficulty, for example, in discriminating between activities with equal risk but with different age patterns of incidence.

Both methods, whether actuarial or based on welfare theory or not, suffer a common major deficiency. They are fundamentally partial-equilibrium approaches. They ignore the chain of wider economic transfers set up through society when life is lengthened. To return to the earlier example, willingness to pay, as currently interpreted, would approve an advance in life from seventy to eighty years if those affected and their kin were willing to pay the cost of the increase. Forgotten, however, is that prolongation of life is not costless to wider society: those who live longer, consume longer, and this extra consumption must be financed by the production of those at younger labor force ages. Proper accounting, we would suspect, should include intergenerational transfer costs, felt in this case as a heavier Social Security burden on the young.

This paper seeks a method that (a) is fully actuarial, (b) is based on welfare theory, and (c) includes economic transfers across society. It sets out to deduce the life cycle implications of changes in the mortality age pattern within a simple general-equilibrium framework with full age-specific accounting.

I. The Economics of Changes in Mortality Risk

To set the context for the analysis, I first set up a neoclassical, age-specific model

*Research scholar, International Institute for Applied Systems Analysis, Laxenburg, Austria. I would like to thank Neil Ericsson and Nathan Keyfitz for valuable discussions on the ideas of this paper, and George Borts, Michael Jones-Lee, Joanne Linnerooth, Samuel Preston, and an anonymous referee for useful comments on an earlier draft.

[1] See, for example, Burton Weisbrod or Simon Rottenberg. Whether earnings should be net of the individual's consumption or not has been the subject of some contention.

[2] See for example E. J. Mishan. Also see the recent works of Bryan Conley, Dan Usher, and M. W. Jones-Lee that put the approach on a quantitative footing by modeling the rational person's willingness to buy extra life years and valuing it in consumption terms.

of the Samuelson consumption-loans type. Within this model of the economy and population, the effect of a change in the mortality pattern on life cycle well-being is then derived.

A. *Neoclassical Model*

Begin with the economy. Output is produced by combining capital K with labor L in a constant-returns production function F.[3] The economy stores no consumption goods. Output is split into consumption and investment in capital growth. Thus

$$(1) \quad F(K(\mathrm{t}), L(\mathrm{t})) = C(\mathrm{t}) + DK(\mathrm{t}) \quad C(\mathrm{t}) \geqq 0$$

The population grows according to the Lotka dynamics which relate this year's flow of births, $B(\mathrm{t})$, to those born in the past by

$$(2) \quad B(\mathrm{t}) = \int_0^\omega B(\mathrm{t}-x)p(\mathrm{t},x)m(\mathrm{t},x)\,dx$$

where $p(\mathrm{t},x)$, the survival function, is the proportion of those born at time $\mathrm{t}-x$ who survive to age x, and $m(\mathrm{t},x)$, the fertility function, is the proportion reproducing at age x, time t; ω is an upper bound on the length of life and the initial birth sequence is assumed given. I assume the same survival or mortality function p is faced by everyone, and it cannot be altered by the individual.

Assume the population is *stable*,[4] and is growing exponentially at rate g. In this case equation (2) has the solution

$$(3) \quad B(\mathrm{t}) = B(0)e^{g\mathrm{t}}$$

where the growth rate g is connected to the age-specific functions p and m by substituting (3) in (2), and canceling B to yield

$$(4) \quad 1 = \int_0^\omega e^{-gx}p(x)m(x)\,dx$$

If $\lambda(x)$ is the age schedule of labor participation, the labor force L and total population N are given by

$$(5) \quad L(\mathrm{t}) = \int_0^\omega B(\mathrm{t}-x)p(x)\lambda(x)\,dx = B(\mathrm{t})\int_0^\omega e^{-gx}p(x)\lambda(x)\,dx$$

$$(6) \quad N(\mathrm{t}) = \int_0^\omega B(\mathrm{t}-x)p(x)\,dx = B(\mathrm{t})\int_0^\omega e^{-gx}p(x)\,dx$$

The labor/population ratio L/N and the birth rate B/N will be denoted by $h(g)$ and $b(g)$, respectively.

Individual consumption varies with age, as do the mortality, fertility, and labor participation rates above. (How it varies is determined below.) Putting population and economic variables together, we can express total consumption C as the sum of individual age-related consumption $c(\mathrm{t},x)$ by

$$(7) \quad C(\mathrm{t}) = \int_0^\omega B(\mathrm{t}-x)p(x)c(\mathrm{t},x)\,dx$$

Later we shall need the average ages of producing A_L, consuming A_c, and reproducing A_m, in the population, defined by

$$A_L = \int_0^\omega xe^{-gx}p(x)\lambda(x)\,dx \Big/ \int_0^\omega e^{-gx}p(x)\lambda(x)\,dx$$

$$A_c = \int_0^\omega xe^{-gx}p(x)c(\mathrm{t},x)\,dx \Big/ \int_0^\omega e^{-gx}p(x)c(\mathrm{t},x)\,dx$$

$$A_m = \int_0^\omega xe^{-gx}p(x)m(x)\,dx \Big/ \int_0^\omega e^{-gx}p(x)m(x)\,dx$$

Assuming the economy has reached a Solow-type steady state, where the growth

[3] F is assumed concave, first-degree homogeneous, and continuously differentiable; for simplicity, capital depreciation is ignored.

[4] That is, its age-specific rates of fertility and mortality, and its normalized age distribution are all constant over time; g is assumed positive.

rate of the economy equals that of population and per capita variables are constant, and assuming society's agent, the government, ensures that investment is maintained at a level that maximizes sustainable total consumption, then

$$(8)\qquad (dK/dt)K=g;\quad I=gK$$
$$c(t,x)=c(x);\quad F_K=g$$

One central fact in society is that consumption, which takes place at all ages, must be supported by production, which takes place only at labor-participative ages. The economy in other words functions at all times under the budget identity $C\equiv F(K,L)-gK$, i.e.,

$$(9)\quad \int_0^{\omega} B(t-x)p(x)c(x)\,dx$$
$$\equiv(F/L-gK/L)\int_0^{\omega} B(t-x)p(x)\lambda(x)\,dx$$

Using (3) and dividing through by $B(t)$, with usual per unit labor notation this societal budget constraint becomes

$$(10)\quad \int_0^{\omega} e^{-gx}p(x)c(x)\,dx$$
$$\equiv(f(k)-gk)\int_0^{\omega} e^{-gx}p(x)\lambda(x)\,dx$$

Thus intergenerational transfers of consumption are introduced by the inescapable requirement that, when growth, labor participation rates, and the capital-labor ratio remain unchanged, any increase in consumption for one age group must be matched by decreases for other age groups.

To complete the model, it remains to determine the life cycle pattern of consumption. Let $U[c,x]$ be the utility rate of being alive at age x, given consumption rate c. The function U is specified up to a multiplicative constant and is concave and continuously differentiable with respect to c. People leave no bequests, they do not fear death, and they individually allocate their consumption to maximize expected lifetime welfare W, where

$$(11)\qquad W=\int_0^{\omega} U[c(x),x]p(x)\,dx$$

In aggregate, of course, they must do this in such a way that the societal budget constraint continues to hold at all times. Standard consumption-loans mechanisims ensure that this happens: a market interest rate and social insurance arrangements appear which encourage people to distribute lifetime consumption to maximize W so that the social budget constraint is always met.[5] Finding the life cycle consumption pattern is thus a simple constrained variational problem, the solution of which yields

$$(12)\qquad \partial U/\partial c(x)=e^{-gx}\partial U/\partial c(0)$$

Thus people pattern their consumption according to age-related need so that its marginal usefulness is the same at all ages, modified only by the ability to invest at an interest rate g, which equals the rate of population growth. Condition (12) therefore is the continuous-age generalization of Paul Samuelson's "biological interest rate" condition.

All preliminaries are now completed. Population and economic growth are well-defined ((3), (4) and (8)), as is the pattern of life cycle consumption ((10) and (12)). And the societal budget identity (10) connects the demography of consumption with that of production.

B. *Change in Age-Specific Risks*

I now introduce a particular, but small, age-specific change in age-specific risks, so that the mortality schedule $p(x)$ becomes $p(x)+\delta p(x)$, and derive the implications for our chosen criterion—the representative

[5] I have in mind here a Menahem Yaari-type social security arrangement where the individual can purchase (or sell) notes at a certain age to be redeemed at a later age, only if he lives, at an actuarially fair return plus the market-clearing interest rate g. If he dies, he and his heirs forfeit all claims.

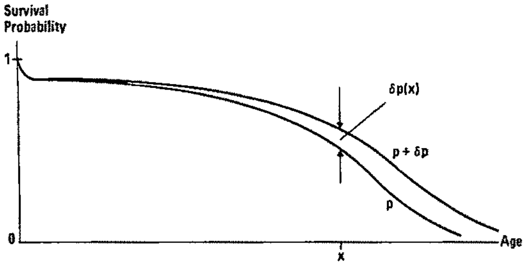


FIGURE 1. CHANGE IN AGE SPECIFIC MORTALITY RISKS
Note: Change shown is for a decrease in the incidence of cancer (scale of δp exaggerated).

person's expected lifetime welfare W.[6] For convenience I assume the mortality variation lengthens life; for shortened life the argument is symmetrical.

When the mortality schedule changes, several variables are forced to change with it: the growth rate g, the consumption pattern $c(x)$, life cycle welfare W, and others. Introducing some new notation, we can write the following expectations: of extra utility from lengthened life U_{ex}; of extra lifetime consumption c_{ex}; of extra man-years of production L_{ex}; and of additional births v_{ex}; due to the particular mortality variation δp.

$$(13) \quad U_{ex} = \int_0^\omega U[c(x),x]\delta p(x)\,dx$$

$$c_{ex} = \int_0^\omega e^{-gx}c(x)\delta p(x)\,dx$$

$$L_{ex} = \int_0^\omega e^{-gx}\lambda(x)\delta p(x)\,dx$$

$$v_{ex} = \int_0^\omega e^{-gx}m(x)\delta p(x)\,dx$$

(The last three are discounted because future consumption utilities will be valued in what follows from date of birth.)

To derive first $\delta g[\delta p]$, the change in the intrinsic growth rate due to the mortality variation, recall equation (4):

$$1 = \int_0^\omega e^{-gx}m(x)p(x)\,dx$$

Using the appropriate chain rule

$$0 = \int_0^\omega e^{-gx}m(x)\delta p(x)\,dx - \delta g\int_0^\omega xe^{-gx}m(x)p(x)\,dx$$

whence[7]

$$(14) \quad \partial g[\delta p] = \frac{\int_0^\omega e^{-gx}m(x)\delta p(x)\,dx}{\int_0^\omega xe^{-gx}m(x)p(x)\,dx} = v_{ex}/A_m$$

The altered mortality pattern affects the growth rate by the change in reproductive value at birth divided by the average age of reproduction (average length between generations). Note that if the mortality variation affects only postreproductive ages, v_{ex} is zero, so that no change in the growth rate occurs.

We can now derive the change in expected lifetime welfare, $\delta W[\delta p]$, from (11) as

$$(15) \quad \delta W = \int_0^\omega U[c(x),x]\delta p(x)\,dx + \int_0^\omega \partial U/\partial c(x)\delta c(x)p(x)\,dx$$

$$= \int_0^\omega U[c(x),x]\delta p(x)\,dx + \partial U/\partial c(0)\int_0^\omega e^{-gx}\delta c(x)p(x)\,dx$$

Life cycle welfare is changed directly by extra years and indirectly by the alteration

[6] A word on the choice of expected lifetime utility as the social criterion. It is quite legitimate to ask what are the consequences of risk change for *any* arbitrary criterion. Suitability of a particular criterion depends on how well it "represents" social interests and on the "reasonableness" of the implications, both judgmental matters. Given individuals who allocate their consumption to maximize expected lifetime well-being, W is arguably representative. Reasonableness of implications will be judged later.

[7] Technically $\delta g[\delta p(x)]$ is a Fréchet differential—a differential whose argument is a function and not a single valued variable.

in the consumption pattern needed to accommodate these extra years. The latter can be evaluated by taking differentials across the societal budget identity (10):

$$
\begin{aligned}
(16)\quad 0= & \int_0^{\omega} e^{-gx} c(x)\delta p(x)\,dx \\
& +\int_0^{\omega} e^{-gx}\delta c(x)p(x)\,dx \\
& -(f(k)-gk)\int_0^{\omega} e^{-gx}\lambda(x)\delta p(x)\,dx \\
& -\delta k(f'-g)\int_0^{\omega} e^{-gx}\lambda(x)p(x)\,dx-\beta\delta g
\end{aligned}
$$

$$
\begin{aligned}
\text{where}\quad \beta= & \int_0^{\omega} xe^{-gx}c(x)p(x)\,dx \\
& -(f(k)-gk)\int_0^{\omega} xe^{-gx}\lambda(x)p(x)\,dx \\
& -k\int_0^{\omega} e^{-gx}\lambda(x)p(x)\,dx
\end{aligned}
$$

From the savings rule $f'=g$ the fourth term in (16) disappears. Where $\bar{c}$ is per capita consumption, β, the life cycle value of a marginal increase in the growth rate, can be expressed as[8]

$$
(17)\qquad \beta=\frac{1}{b}\left[\bar{c}(A_c-A_L)-kh\right]
$$

Finally, using (16) to substitute for the second term in (15), and noting that for constant returns $f-kg$ is F_L, we obtain

$$
\begin{aligned}
\delta W= & \int_0^{\omega} U[c(x),x]\delta p(x)\,dx \\
& +\frac{\partial U}{\partial c(0)}\left\{F_L\int_0^{\omega} e^{-gx}\lambda(x)\delta p(x)\,dx\right. \\
& \left.-\int_0^{\omega} e^{-gx}c(x)\delta p(x)\,dx+\beta\delta g\right\}
\end{aligned}
$$

Re-expressed in more convenient notation this becomes my first main result. The net life cycle utility value of a particular age-specific change in mortality risk is given by[9]

$$
(18)\qquad \delta W=U_{ex}+\frac{\partial U}{\partial c(0)}\{F_L L_{ex}-c_{ex}+v_{ex}\beta/A_m\}
$$

This result is easily interpreted. When mortality is improved, the individual is blessed with extra expected years of life, extra expected years of productive work if retirement years are affected, and extra expected children if reproductive years are affected. These, valued in utility units, are the first, second, and fourth terms on the right of (18). On the other hand, extra years must somehow be supported. The third term shows the total amount of consumption support needed—either a burden on Social Security and hence consumption at earlier ages, or a burden directly on one's own production, depending on whether post- or preretirement years are affected.

These welfare changes will occur in general at different periods in the life cycle. Those in the productive age groups tend to carry the consumption cost; usually only in later life do they reap the utility of extra years, with the costs turned over to a new generation. To the extent that population is

[8] For a discussion of β, the life cycle value of a marginal increase in the growth rate, see my paper with Geoffrey McNicoll.

[9] It would be possible to complicate these results in various ways. Where increased longevity induces a shift in the labor-participation schedule or in the retirement age, L_{ex} should be expanded to include increased participation, as well as increased survival. Where person i's utility rate includes the enjoyment α_j^i that loved ones j (with age differences a_j) are alive, life cycle welfare becomes

$$
W^i=\int_0^{\omega}\left[U^i+\sum_j \alpha_j^i p(x+a_j)\right]p(x)\,dx
$$

Here (18) would contain an extra kith-and-kin term:

$$
\sum_j \alpha_j^i\int_0^{\omega}(\delta p(x+a_j)p(x)+p(x+a_j)\delta p(x))\,dx
$$

Increased survival is therefore twice valuable—for any person it increases the chances his parents and grandparents will survive to be enjoyed, and the chances that he will survive to enjoy his children and grandchildren.

growing, younger age groups are larger than older ones and transfers toward later times and ages are easier on the individual; this is why the analysis discounts costs at rate g over the life cycle in the above terms.

II. Value of Life

Until now I have viewed activities that put life under hazard in rather inconvenient terms as causing variations in the mortality age profile. Is it possible to proceed more directly and value actual lives lost or saved? In the literature, most writers prefer to deal with marginal changes in risk rather than with direct loss of life, feeling possibly that increase of risk is more approachable and somehow less awesome than loss of life. From an actuarial viewpoint, however, risk and death cannot be separated. For any sizeable population, an increase in age-specific risk means, in life table terms, an increase in numbers of deaths at specific ages. We might therefore expect valuation of risk and valuation of lives lost to be closely connected.

Let us approach the valuation of lives lost by asking a specific question. Suppose in the community an unspecified activity were to take one life at random at age a, how much welfare would the community as a whole be prepared to give up to rid itself of the increased risk?[10] The result will be called the social welfare equivalent (SWE) of life, at age a.

To answer this question we may recall that $p(x)$, the life table function, is calculated by taking a base number of births $\bar{B}$ (for example 10,000) and observing the year-by-year decrements in survivorship. Assume now that every $\bar{B}$ people born undergo one additional death at age a. Until age a there is no difference in survivorship; at age a there are $\bar{B}p(a)-1$ survivors instead of $\bar{B}p(a)$; at age $x>a$ there are $(\bar{B}p(a)-1)p(x)/p(a)$ survivors instead of $\bar{B}p(x)$. The additional death therefore causes a variation in the mortality schedule (see Figure 2) equal to the difference in numbers surviving divided by the base:

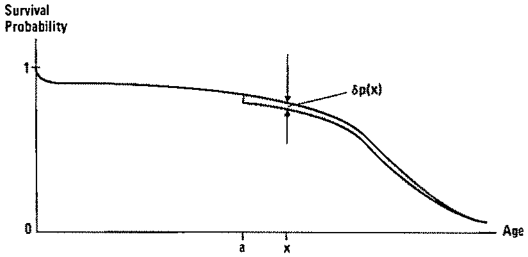


FIGURE 2. MORTALITY VARIATION CAUSED BY A SINGLE ADDITIONAL DEATH AT AGE a

Note: Scale of δp exaggerated.

$$(19)\quad \delta p(x)=\begin{cases}0 & 0\leqq x\leqq a\\ p(x)/p(a)\bar{B} & a<x\leqq\omega\end{cases}$$

I shall write $p(x)/p(a)$ as $p_a(x)$, the probability of survival to age x given survival already to age a.

I have now translated the value-of-life problem into one of valuing changes or variations in the mortality schedule; hence we can use the machinery of the previous section. Substituting the variation (19) into equation (18), the additional death imposes a risk that lowers the expected life cycle welfare of each representative individual an amount

$$(20)\quad WE=\int_a^\omega U[c(x)]\frac{p_a(x)}{\bar{B}}dx+\frac{\partial U}{\partial c(0)}\times\left\{F_L\int_a^\omega e^{-gx}\lambda(x)\frac{p_a(x)}{\bar{B}}dx-\int_a^\omega e^{-gx}c(x)\frac{p_a(x)}{\bar{B}}dx+\frac{\beta}{A_m}\int_a^\omega e^{-gx}m(x)\frac{p_a(x)}{\bar{B}}dx\right\}$$

It would take $\bar{B}$ times this amount to compensate the total number of persons at risk, $\bar{B}$. Hence we multiply (20) by $\bar{B}$ to arrive at

[10] Within the assumptions made earlier, it is legitimate to evaluate a certain but random death by the difference this risk makes to the representative person's expected well-being. Not all writers would agree with this procedure in general though: see John Broome, for example.

the *SWE* that would compensate the community for the small increased risk corresponding to loss of one life at age a. This yields our second main result—a result that has an obvious actuarial interpretation.

$$(21)\quad SWE=\int_a^\omega U[c(x)]\,p_a(x)\,dx + \frac{\partial U}{\partial c(0)}\int_a^\omega e^{-gx}\left\{F_L\lambda(x)p_a(x) - c(x)p_a(x)+\frac{\beta}{A_m}m(x)p_a(x)\right\}dx$$

Equation (21) shows that the social welfare equivalent of a loss of life aged a equals the value of remaining expected years of life at age a, remaining expected labor years at age a, and remaining expected reproduction at age a, minus the cost of remaining expected consumption upkeep at age a.

Assume the utility and consumption rates are roughly constant at $U(a)$ and $c(a)$ over the remaining years; where $w(=F_L)$ is the wage rate; where e_x denotes the expected value of remaining survival years at age x, and where $\tilde{e}_x$, $\tilde{e}_{lx}$, $\tilde{e}_{mx}$ are the discounted expected values of remaining survival years, labor years, and net fertility at age x, we can write (21) in the useful form

$$(22)\quad SWE(a)=U(a)e_a + \frac{\partial U}{\partial c(0)}\left\{w\tilde{e}_{la} - c(a)\tilde{e}_a + \frac{\beta}{A_m}\tilde{e}_{ma}\right\}$$

A marginal life lost is therefore valued in terms of opportunity lost—opportunity to enjoy further life, to produce further output, to have additional children, less, of course, consumption support costs that are no longer necessary.

To extend the analysis to the case of numbers of lives lost at various ages, consider an activity R (air travel, say) that costs De^{gt} lives in year t, where the numbers of deaths are small relative to total deaths. Assume these deaths are distributed as $d(a)e^{gt}$ at age a, so that the probability that a life lost to this activity is age a is $\phi_R(a)=d(a)/D$. In our analysis the cost of lives lost is imputed to this year's cohort which stands to lose $d(a)e^{g(t+a)}$ lives at age a in year $t+a$. The value-of-life argument above is additive over lives lost, therefore for this activity in year t, total (welfare-equivalent) losses are

$$\text{Total } SWE=\sum_a d(a)e^{gt}e^{ga}SWE(a)$$

Finally, multiplying above and below by D, gives the needed result

$$(23)\quad \text{Total } SWE=De^{gt}\sum_a \phi_R(a)e^{ga}SWE(a)$$

Cost of lives lost, in other words, is the number of deaths per year times the expected cost of a death in the activity in question.[11]

III. Discussion and Illustrations

A. *Robustness*

In the foregoing analysis I assumed an idealized world of economic and demographic steady-state growth, constant returns in production, perfect life cycle financial markets, and similar individuals who face similar mortality schedules. How robust are the results when these assumptions are replaced by more realistic ones?

Note first that the important factors are changed but little under increased realism. When risks to life fall for the population (a) the individual does enjoy expected extra years, extra expected working life, and extra expected reproduction exactly as above, and (b) whatever the support mechanism for old age, be it gifts to tribal elders, child support for ageing parents, or an *ad hoc* government Social Security system, consumption must

[11] The e^{ga} factor enters to preserve consistency: the cost-of-loss-of-life argument was developed on a cohort (life cycle) basis, whereas deaths are introduced on a period (current year) basis.

still be set aside for lengthened life (although the amount depends on the transfer mechanism, and the analysis of these cases requires changes in the model.) With nonconstant returns in production and imperfect life cycle markets, the valuation of these factors changes however. The marginal value of consumption may well vary more widely than in (12), labor may not necessarily be paid its marginal product, and the value of growth β will be altered. With nonoptimal investment, an extra capital-labor ratio adjustment term enters.

These changes are minor, however, compared to the case where altered mortality risks strike the population unevenly, or the mortality change comes suddenly, or demographic and economic growth vary widely from steady state. In this case, some people may reap the benefits of increased life and production, while others bear the consumption costs. For example, a sudden mortality improvement can be a windfall to the elderly—they enjoy extra years while escaping the corresponding extra support of the generation that went before.

B. *Additional Living versus Additional Consumption*

Any risk evaluation method must unavoidably compare two very different things; the enjoyment of additional living U, and the enjoyment of additional consumption $\partial U/\partial c$. To make this comparison more explicit, for the rest of the paper I shall narrow the results to the special case where the form of the utility function U does not vary with age, and U has constant elasticity of consumption, $\varepsilon(=(dU/dc)(c/U(c)))$. The parameter ε will serve as a useful proxy for the tradeoff between utility of living and utility of consumption. When it is low (near zero), proportional increases in consumption have little effect and we can think of utility as derived almost solely from the state of being alive.[12] When it is relatively high (near one), U becomes linear in consumption; we could therefore recalibrate U in consumption terms—here, in a sense, utility *is* consumption. Normally, we will allow ε to take an intermediate value.

Under this special form of U, some further algebra shows that (18) reduces to

(24)
$$\delta W=\frac{\partial U}{\partial c(0)}\left[\left(\frac{1}{\varepsilon}-1\right)c_{ex}+wL_{ex}+\frac{\beta}{A_m}v_{ex}\right]$$

where $w(=F_L)$ is the wage rate. Utility of additional years, U_{ex}, now translates directly into consumption terms as c_{ex}/ε. Dropping the $\partial U/\partial c(0)$ factor, we may express the value of the mortality change to the individual directly as marginal consumption equivalent to

(25)
$$CE[\delta p]=\left(\frac{1}{\varepsilon}-1\right)c_{ex}+wL_{ex}+\frac{\beta}{A_m}v_{ex}$$

As we would expect then, a crucial but arbitrary element in the evaluation of mortality change is the degree to which pure enjoyment of additional years is offset by its consumption cost. In our well-off society we could expect additions to longevity to outweigh consumption considerations (ε is low), but in poorer societies (ε is high) utility of additional living might be offset by the additional burden of support; in certain nomadic societies for example, older members, if no longer productive, are expected to separate from the tribe and die.[13]

One often hears two different ethical arguments where activities that put life at risk are under discussion: "life is infinitely valuable" vs. "social product is what counts." As we would again expect, these follow from different positions on the living/consumption tradeoff. When ε tends to zero, (25) shows that additional life years

[12] Where U has the constant elasticity form $U=\alpha\cdot c^{\varepsilon}$, it tends to the constant level α over all c as ε tends to zero. Ruled out is the case $\varepsilon>1$, where U ceases to be concave in c violating the assumptions.

[13] Even in Western society, life could not be extended much beyond 100 years unless retirement age were also increased. See Kenneth Boulding for an entertaining essay on the economic menace of extreme longevity.

TABLE 1–AGE-SPECIFIC SURVIVAL SCHEDULES AND VARIATION CAUSED BY ELIMINATION OF CARDIOVASCULAR DISEASES[a]

Age	0	10	20	30	40	50	60	70	80
$p_E(x)$	1.000	.97029	.96402	.95260	.93747	.91056	.85807	.76547	.62615
$p(x)$	1.000	.97000	.96343	.95091	.93149	.88841	.79228	.61135	.34853
$\delta p(x)$	0.000	.00029	.00059	.00169	.00598	.02215	.06579	.15412	.27762

[a]From latest available cause-deleted lifetables: S. H. Preston et al. for *U.S.* 1964. Mortality in the United States has changed but little in the last sixteen years. See Preston for further details on cause of death.

outweigh any consumption considerations: activities should be judged only on whether they preserve and prolong life. When ε becomes one, extensions to life are perfectly offset by their consumption cost: only social product considerations remain. Normally valuation of mortality change would retain elements of both ethical positions.

Equation (25) can be used to comment on the two methods in present use. *Willingness to pay*, as usually interpreted, ignores the negative social burden term.[14] It will therefore tend to overstate the value of mortality reduction and unduly bias against risky projects. *Human capital* tends to understate and therefore to bias toward risky projects. Only in the special case where (a) altered risks do not affect childbearing ages, (b) population growth is vanishingly small, and (*c*) utility shows constant returns to consumption ($\varepsilon=1$), would the (gross) human capital method be justifiable and correct. In this case, additional life years would be exactly offset by their consumption cost, so that (25) would reduce to the human capital measure $CE=wL_{ex}$.

C. *An Example: Cardiovascular Diseases*

To illustrate (25), let us assess the worth to the individual of elimination of cardiovascular diseases in the United States. Table 1 shows p, the standard survival function, and p_E, the survival function if cardiovascular diseases were eliminated. The difference between them is δp.

Under 1975 *U.S.* data[15] and the definitions in (13), complete elimination of cardiovascular diseases yields the following differentials:

$$\text{Extra Years}=7.69; \quad c_{ex}(\$)=42{,}670;$$

$$L_{ex}(\text{years})=0.692; \quad v_{ex}=0.00135.$$

Cardiovascular diseases attack for the most part postproductive and postreproductive age groups. Hence, though longevity increases significantly, expected working life and expected number of children increase only a little.

Where $\varepsilon=1.0$, 0.6, and 0.4, from (25) we obtain

$$C.E.=\left.\begin{matrix}(1.0-1)\,42{,}670\\(1.667-1)\,42{,}670\\(2.5-1)\,42{,}670\end{matrix}\right\}$$

$$+(13{,}749)0.692+(-68{,}125)0.00135$$

$$=\begin{cases}\$9{,}400\\ \$37{,}800\\ \$73{,}400\end{cases}$$

This of course does not imply the United States should spend corresponding amounts per person on the elimination of cardiovascular diseases. A flood of research dollars would by no means guarantee such a breakthrough. The illustration, however, gives an idea of the potential returns to the individual.

[14]Not always. See Conley, for example, and Jones-Lee (1976, p. 44, fn. 42).

[15]The illustrations for *U.S.* 1975 are for male and female combined. Source for labor participation schedule was ILO Year Book 1976, for fertility schedule was *Statistical Abstract of the United States* 1977. The wage rate was computed as $13,749 and β as –$68,125. For further details of data and computations see my 1979 paper.

TABLE 2—EXPECTED ADDITIONAL LIFE YEARS, LABOR YEARS, AND REPRODUCTION, AND ILLUSTRATIVE COST OF LOSS OF LIFE AT AGE *a*

Age *a*	0	10	20	30	40	50	60	70	80
$\tilde{e}_a$ (years)	70.3	62.5	52.9	43.5	34.3	25.6	18.0	11.7	6.7
$\tilde{e}_{la}$ (years)	31.6	32.5	31.4	24.7	17.6	10.8	4.4	0.3	–
$\tilde{e}_{ma}$	0.921	0.949	0.882	0.339	0.038	–	–	–	–
SCE ($1,000)									
$\varepsilon=1.0$	371	382	371	316	239	148	61	4	–
$\varepsilon=0.6$	668	664	619	520	399	265	139	54	31
$\varepsilon=0.4$	1,055	1,031	942	783	605	417	241	119	72

Notes: Dollar figures are 1975 dollars. These figures do not include any cost to kin of the loss of life of their loved one. *SCE* at age 0 would not be a suitable way to measure the desirability of introducing an additional birth: the analysis calculates how much those *already* born would give up to avoid certain types of risk.

TABLE 3—AGE PATTERNS OF INCIDENCE FOR THREE CAUSES OF DEATH

Age	0	10	20	30	40	50	60	70	80
Motor vehicle	.027	.045	.224	.152	.119	.117	.119	.116	.081
Maternal	–	–	.25	.452	.299	–	–	–	–
Neoplasms	.003	.005	.006	.014	.041	.116	.229	.319	.267

Source: Preston et al.

TABLE 4—COMPARISON OF PREVENTING DEATH FROM THREE ALTERNATIVE CAUSES

	Expected Additional			*SCE* $1,000 (1975)		
	Survival Years	Labor Years	Net Reproduction	$\varepsilon=1.0$	$\varepsilon=0.6$	$\varepsilon=0.4$
Motor Vehicle	34.3	17.0	0.32	212	369	574
Maternal	43.1	24.3	0.39	307	509	770
Neoplasms	15.5	3.9	0.02	52	121	221

D. *Value of Life*

In the special constant elasticity case, as in the previous subsection, we can re-express the value-of-life expression (22) as a social consumption equivalent (*SCE*) of a life at age *a*:

(26)

$$SCE(a)=\left(\frac{1}{\varepsilon}-1\right)c(a)\tilde{e}_a+w\tilde{e}_{la}+\frac{\beta}{A_m}\tilde{e}_{ma}$$

Table 2 gives an idea of the magnitude of the *SCE* at different ages and different elasticity values, and shows that the cost of a life lost, under the chosen criterion of expected lifetime welfare, is highly age dependent. We might therefore expect the average cost of loss of life to be different for activities that strike different age groups. We can compare the gain to saving (restoring to normal survival probabilities) a life chosen randomly, otherwise lost to motor vehicle accident death, maternal death, or cancer death. Table 3 shows probability distributions over age, $\phi(a)$, for deaths due to these causes.

The expected gain in saving one life at random in year 0 follows from (23) as

$$SCE=\sum_a \phi_R(a)e^{ga}SCE(a) \qquad (27)$$

From this expression, and the above tables we obtain Table 4. It should be noted that the effort or cost required to prevent loss of life may be quite different in each of these causes and is not considered here.

IV. Conclusion

This paper derived expressions for the value of activities that alter the mortality schedule and for the cost of premature loss of life, under specific assumptions and a life cycle welfare criterion. A change in the pattern of the mortality schedule, it was shown, should be assessed by the difference it makes to expected length of life, production, reproduction, and consumption support; loss of life should be assessed by the expected opportunity costs of lost years, production and reproduction, less support costs.

Valuation of risks to life depends heavily on age, as the illustrations above show. This is a consequence of the chosen life cycle welfare criterion, under which a life lost at a younger age forfeits more than one at an older age. If it were felt, on the other hand, that "a life is a life whatever the age," then a life cycle criterion would no longer be appropriate.

Social support costs figure large in the valuation of risks to life. The degree to which these offset the pure enjoyment of staying alive can make a significant difference to numerical assessments. Where being alive is valued much more highly than pure consumption, additional support costs, like additional wage earnings, fade from significance. But where the value of being alive is measured purely by additional consumption, the gain from added longevity can be cancelled completely by the additional consumption support required.

REFERENCES

J. P. Acton, "The Value of Life: An Overview and Critique of Alternative Measures and Measurement Techniques," *Law Contemp. Probl.*, Autumn 1976, *40*, 46–72.

W. B. Arthur, "The Economics of Risks to Life," RR 79-16, Int. Inst. Appl. Systems Anal., Dec. 1979.

______ **and G. McNicoll,** "Samuelson, Population, and Intergenerational Transfers," *Int. Econ. Rev.*, Feb. 1978, *19*, 241–46.

K. E. Boulding, "The Menace of Methuselah: Possible Consequences of Increased Life Expectancy," *J. Washington Acad. Sci.*, Oct. 1965, *55*, 171–79.

J. Broome, "Trying to Value a Life," *J. Publ. Econ.*, Feb. 1978, *9*, 91–100.

B. C. Conley, "The Value of Human Life in the Demand for Safety," *Amer. Econ. Rev.*, Mar. 1976, *66*, 45–55.

M. W. Jones-Lee, "The Value of Changes in the Probability of Death or Injury," *J. Polit. Econ.*, July/Aug. 1974, *82*, 835–49.

______, *The Value of Life: An Economic Analysis*, London 1976.

J. Linnerooth, "The Value of Human Life: A Review of the Models," *Econ. Inquiry*, Jan. 1979, *17*, 52–74.

E. J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," *J. Polit. Econ.*, July/Aug. 1971, *79*, 687–705.

S. H. Preston, *Mortality Patterns in National Populations*, New York 1976.

______, **N. Keyfitz, and R. Schoen,** *Causes of Death: Life Tables for National Populations*, New York 1972.

S. Rottenberg, "Economics of Health: The Allocation of Bio-Medical Research," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 109–18.

P. A. Samuelson, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *J. Polit. Econ.*, Dec. 1958, *66*, 467–82.

D. Usher, "An Imputation to the Measure of Economic Growth for Changes in Life Expectancy," in Milton Moss, ed., *The Measurement of Economic and Social Performance*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 38, New York 1973.

Burton A. Weisbrod, *Economics of Public Health*, Philadelphia 1961.

M. E. Yaari, "Uncertain Lifetime, Life Insurance and the Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, *32*, 137–50.

U.S. Bureau of the Census, *Statistical Abstract of the United States* 1977, 98th ed., Washington 1977.

Two-Part Tariffs and Optimum Taxation: The Case of Railway Rates

By Sylvester Damus*

Suppose railways have increasing returns to scale, or are burdened with substantial common costs that create a condition resembling increasing returns. Average cost would then exceed marginal cost, and unsubsidized private operation would require a divergence of price from marginal cost which is a hindrance to internal commerce and a burden on society. A. C. Pigou and Harold Hotelling would have solved this problem by a combination of taxes and subsidies, whereas Ronald Coase suggested that two-part tariffs be used instead. His idea of price as a fee for service plus a tax can be traced back to Arthur Hadley (p. 137, fn. 1) and earlier writers. The tax in Coase's tariff would be assessed on the commodities carried by railway. Frank Ramsey set the rules for optimum taxation which then ought to be followed. His more general rule is that taxes should bring about an equiproportional reduction of all units of service from their marginal cost-pricing levels. By another version of his rules, the tax rate should be in inverse proportion to elasticity of demand. This rule is not very different from the practice of a profit-maximizing monopolist, who also raises price above marginal cost in inverse proportion to elasticity of demand.

Economists can only discover what the industry has known for quite some time. Indeed, William Baumol and David Bradford traced the roots of Ramsey pricing back to E. P. Alexander, an officer of the Louisville and Nashville Railroad, and director of the Union Pacific. In this paper I test the hypothesis that railways set Ramsey prices whenever they could. I say "whenever they could," because Ramsey prices were not sustainable against direct competition.

To test this hypothesis the hard way, one estimates demands and marginal costs of multiproduct firms, collects data on prices, finds the "Ramsey number" for each product line—the percentage deviation of price from marginal cost times the elasticity of demand—and tests for the equality of Ramsey numbers across product lines. In addition, one would have to verify that firms break even by this method of pricing.

In the case of railways, an easier approach can be followed. Let us define the multiple products as short, intermediate, and long hauls, and find them priced according to a distance tariff

$$f_i(x) = t_i + r_i x$$

where $f(x)$ is the freight per ton, t is a *terminal charge* per ton, r is a *conveyance rate* per ton per mile, x is the length of haul in miles, and the subindex i denotes the commodity or class. It will be shown in Section I that this is a two-part tariff set on the tax-cum-subsidy plan: it is equal to marginal cost, plus a tax on tonnage, less a mileage subsidy. This is a new interpretation of tapered distance tariffs, one which implies that t and r are negatively correlated. The hypothesis of negative correlation between t and r is tested in Section II against two rivals: t and r are competitive prices of services rendered per unit of weight and distance; or they are simple monopoly prices of separate services.

Let us next consider the problem when there is an untaxed sector. If the tax is to raise a fixed sum, tax rates can be lowered

[*] Economic Council of Canada. This paper was written while I was on leave from the University of Winnipeg. This paper is based on parts of my dissertation, which had the benefit of supervision by A. C. Harberger, D. N. McCloskey, L. G. Telser, and L. Sjaastad. An anonymous referee and George Borts made substantial improvements to this paper, but are absolved from any further responsibility for its contents. Financial assistance from the University of Winnipeg is gratefully acknowledged.

by broadening the tax base. In the case of railways, the tax base could be broadened by setting up a cartel and reducing competition for long-haul freight. A positive correlation between rates of tax on local traffic and the relative size of the lightly taxed (competitive long-haul freight) sector indicates a constraint on railway profits reaped in local markets. It indicates there was unused room to raise prices towards profit-maximizing levels, whereas profit maximization would imply a zero correlation between tax rates and the size of the competitive sector. In this way we can test the presence of a constraint on the revenue raised by two-part tariffs. It is shown in Section III that this constraint existed well before any effective regulation, although we cannot say how hard it pressed on the firms.

The analysis relies heavily on the work of Charles Ellet and Wilhelm Launhardt. Their theories of profit-maximizing railway rates are similar to Alan Walters' economics of road-user charges, only Walters came to third best conclusions by not considering a vertical integration of the road owner with the road user, such as that accomplished by railways. The theory presented here is an improvement over Gary Hawke's analysis of terminal charges and supports his judgment of the beneficial effects of discriminatory rates. Our evidence of cross subsidization of long-haul shippers by short-haul shippers via a tax-cum-subsidy tariff, and of constraints on the taxation of local traffic by unregulated railways confirms George Stigler's and Claire Friedland's findings of little which regulators can regulate. The evidence of unregulated cross subsidization also has relevance for Richard Posner's theory of taxation by regulation, and Sam Peltzman's theory of regulation.

In the three sections below I present a modified version of Launhardt's model, test the hypothesis that railway distance rates involve cross subsidization of long hauls by short hauls, and test the hypothesis that the net tax in the two-part tariff raised a less than profit-maximizing sum. Both unregulated and regulated rates pass these tests with fairly good marks. It is left to the reader to draw his own conclusions about regulation, and the wisdom of advocating competition per se, rather than for efficiency reasons.

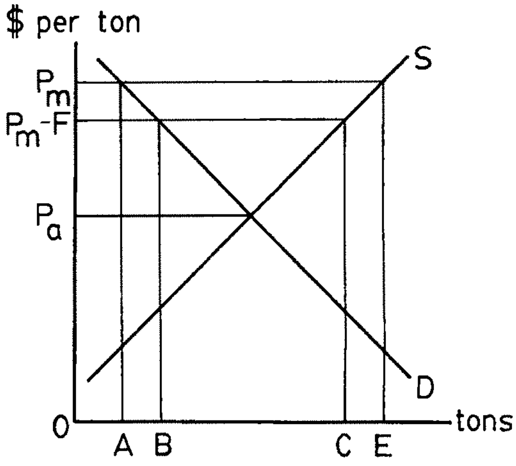


Figure 1. Demand and Supply of a Good at Some Point

I. A Model of Railway Rates

Figure 1 shows the demand and supply of a good at some point on a line of railroad x miles from a market M. The autarchy price of the good at that point is P_a, and P_m is the price it fetches at M. The freight rate for transport to M is some function of distance, $F=f(x)$. If transportation were free, AE tons of the good would be shipped to market. This is the quantity g in Figure 2, the demand for transport. But at the given freight rate for that distance, only BC tons will be shipped, or the amount q in Figure 2. The vertical intercept of the demand for transport v is equal to the difference between P_m and P_a. One may call it the "value of service," in the sense of the freight rate that the traffic will not bear. The value of P_m is of course the same at all points on the line. Assuming that P_a is also the same at every point, the value of service would be constant along the line, and the area under the demand curve for transport, above F, is the locational rent of land. The incidence of the "tax" in the railroad tariff is on this rent. The second best tariff is one that maximizes this rent, subject to a minimum revenue constraint.

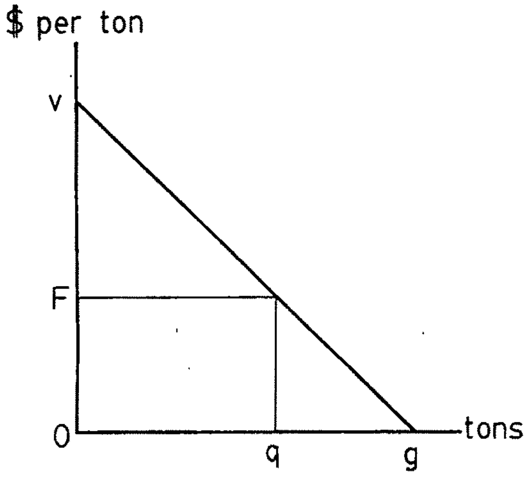


FIGURE 2. DEMAND FOR TRANSPORT AT A POINT

For the backhaul from M, one draws the demand and supply for the backhauled commodity so the autarchy price at x exceeds the price at M. It seems best to assume that the demands for transportation in different directions are independent of each other. Likewise, it will simplify things to assume that the demand for the transportation of a good from one point on a line is independent of the freight on the same good from any other point. This assumption makes it possible to add up the locational rents at all points on the line to obtain the public utility of the road. The length over which rents are added is a miles, up to where $f(a)=v$.

The demand need not be the straight line in Figure 2. I adopt instead the form suggested by Launhardt:

$$(1) \qquad q=g\left[\frac{v-f(x)}{v}\right]^n \qquad n>0$$

The quantity demanded q is determined by shippers who take the railway's tariff schedule of distance rates $f(x)$ as given. Given the length of haul demanded at the shipper's locations, $f(x)$ is the price or rate.

The demand curve is concave from below if n is less than unity, a straight line if n equals unity, and convex if n exceeds unity. The constants v and g are the price and quantity axis intercepts of the demand curve. The maximum possible tonnage g could be made a function of some other variable, say x, without affecting the form and properties of the tariff $f(x)$. It would only affect the level of railway profit, and the level of a tariff determined with a constraint on profit. A $g(x)$ can be used to represent the distribution of traffic between long and short hauls, and the effect of variations in $g(x)$ on tariffs via a profit constraint is the basis for the argument and tests in Section III. The v follows from the small-country assumption made about a region which employs the railway to trade at the price set in a distant market. The power n is a constant elasticity of demand with respect to the fraction of the value of service v conceded by the railway to the shipper. The price elasticity of demand is

$$(2) \qquad E=\frac{-nf(x)}{v-f(x)}$$

The demand is normal, since its price elasticity is zero when the freight rate $f(x)$ is zero, and it tends to infinity as the freight rate approaches v and becomes prohibitive. This is the basis for discrimination between lengths of haul by a distance tariff. Discrimination among commodities is caused by the fact that, given $f(x)$, an increase in v per ton of goods reduces the price elasticity of demand. This normal form is preferred to a constant price elasticity of demand for transport, since the latter would imply that the difference between market prices of commmodities can be as large as one wishes, and thus fail to put a limit on market areas.

The benefit of transport to all shippers of a good located at any point on the line is obtained by integrating the area under the demand curve and above the freight rate. A second integration adds the benefits at all points on the line. The integration is made first between the prices $y_1=f(x)$ and $y_2=v$, and then—assuming that access to the line can be had at any point on it—from zero to a miles from the market M. The upper limit of the second integration is the maximum economically possible length of haul, for which $v=f(a)$ and $q=O$. For any given commodity, the benefit function is

$$(3)\quad g\int_0^a\int_{f(x)}^v\left[\frac{v-y}{v}\right]^n dy\,dx$$
$$=\frac{gv}{n+1}\int_0^a\left[\frac{v-f(x)}{v}\right]^{n+1}dx$$

The marginal cost of carrying a ton of goods from Here to There may be represented by c_i+b_ix, where c_i is the marginal cost of dealing with an additional ton, and b_ix is the total cost of the ton miles of locomotive work put in to take the marginal ton from Here to There. There are also large fixed costs in the amount of K dollars per period of time, independent of both tonnage and ton miles. They may be thought of as the annualized cost of a lumpy factor, the cost of entry by a potential competitor, or as the sum of unallocated common costs.

The similarity in mathematical form between the marginal cost c_i+b_ix and the distance tariff t_i+r_ix has been a source of confusion. It has spawned innumerable attempts to explain away the discrimination in two-part tariffs by reference to costs. The terminal cost of c dollars per ton is here assumed to be zero, to show that the terminal charge t does not necessarily depend on it. It is further assumed that the marginal cost of carrying any commodity is independent of the quantities of other commodities carried. This and the assumed independence of demands allow us to deal with only the "representative commodity."[1] I also assume the marginal cost per ton mile b is independent of distance. To make it a function of distance would change the mathematical form but not the economic substance of tariffs.

The surplus of revenue over and above direct cost equals the quantity times the difference between price and direct cost:

$$(4)\quad S=g\int_0^a\left[\frac{v-f(x)}{v}\right]^n[f(x)-bx]\,dx$$

Maximization of this surplus is the simplest problem in the calculus of variations, because the partial derivative $\partial\phi/\partial f'$ in the Euler equation

$$\frac{\partial\phi}{\partial f}-\frac{d}{dx}\frac{\partial\phi}{\partial f'}=0$$

does not appear in $S=\int_0^a\phi[x,f(x)]\,dx$. Thus we can treat $f(x)$ as an ordinary variable, and look upon results at one point of the line as representative for the whole line, so long as we do not fully distribute the fixed cost K over those points.

At any point, for a given length of haul $\bar{x}$, the freight rate is $f(\bar{x})=F$, and the direct cost is $b\bar{x}=B$. The operating surplus is equal to

$$\frac{g}{v^n}(v-F)^n(F-B)$$

This surplus is maximized by setting the tariff at

$$F=\frac{1}{n+1}v+\frac{n}{n+1}B$$

as is found by differentiation with respect to F. The second derivative is negative, and the surplus can be maximized by this tariff, if the value of service exceeds its price and cost ($v>F>B$). Therefore, the surplus revenue is a concave function of the tariff, and the deficit $K-S$ is convex. Similarly, the benefit to shippers (the expression in equation (3)) evaluated at the profit-maximizing tariff is positive when $v>B$, or at all locations worth a transport service. Its first derivative is negative at all points served, so the benefit falls to zero at $v=F>B$.

To find the second best tariff, let us maximize the concave benefit to consumers subject to a convex budget constraint. This constraint is that marginal cost pricing must avoid deficits: Maximize the expression in (3) with respect to $f(x)$, subject to

$$K-g\int_0^a\left[\frac{v-f(x)}{v}\right]^n[f(x)-bx]\,dx\leqslant 0$$

[1] By dealing with the representative commodity, we shall deemphasize the better known discrimination among commodities. This is carried along by the model, but is pushed into the background to facilitate the exposition of other properties of railway rates.

The Lagrangian function is

$$V=\frac{g}{v^n}\int_0^a\left\{\frac{[v-f(x)]^{n+1}}{n+1}+\lambda[v-f(x)]^n[f(x)-bx]\right\}dx-\lambda K$$

The multiplier λ is the marginal social value of a dollar of railway surplus equal to the opportunity cost of a dollar's worth of deficit finance. It exceeds one dollar, because of the deadweight loss of taxation, assuming that elsewhere there are no distortions. The conditions for a maximum are

$$\frac{\partial\phi}{\partial f}\leqslant 0,\quad f(x)\geqslant 0,\quad f(x)\frac{\partial\phi}{\partial f}=0;$$

$$\frac{\partial V}{\partial\lambda}\geqslant 0,\quad \lambda\geqslant 0,\quad \lambda\frac{\partial V}{\partial\lambda}=0$$

Unsubsidized operation is possible only if $f(x)$ is greater than zero. This and the complementary slackness condition imply

$$\frac{\partial\phi}{\partial f}=0=(\lambda-1)[v-f(x)]-\lambda n[f(x)-bx]$$

whence

$$(5)\quad f(x)=\frac{m}{m+n}v+\frac{n}{m+n}bx,\quad m=1-\frac{1}{\lambda}$$

$$\text{or}\quad f(x)=t+rx;\quad t=mv/(m+n),\quad r=nb/(m+n)$$

Substituting (5) into the equation for surplus over direct cost (4) we have

$$(6)\qquad S=\frac{mn^ngv^2}{(m+n)^{n+1}(n+2)b}$$

The conditions $K\geqslant 0$, $K-S\leqslant 0$, and equation (6) imply in turn $\lambda\geqslant 1$ and $0\leqslant m<1$. Therefore, by the second complementary slackness condition, $\partial V/\partial\lambda=S-K=0$, the budget is balanced exactly.

Since m can vary between zero and unity, there is a continuum of second best tariffs between the limits

$f(x)=bx$, for $m=0$, marginal cost pricing

and $=\frac{1}{n+1}v+\frac{n}{n+1}bx$ for $m=1$, profit maximizing

In the profit-maximizing case, λ tends to infinity, only the railway's surplus has value, and consumer's surplus is regarded as worthless. The choice of one tariff over another depends on K, b, g, n, and v. Hence $K=0$ is the condition for marginal cost pricing. For a poor railway, running through an area in which all parameters assume unfavorable values, the second best tariff may have to be made with $m=1$, the profit-maximizing level. Should the maximum possible surplus over direct cost fall short of K, the line will not be built; or if built, it will not be renewed.

From equations (5) and (2), we find the second best tax rate

$$\frac{f(x)-bx}{f(x)}=-\frac{m}{E}$$

This is one of the formulations of Ramsey's rules for optimum taxation: Prices are set as if the monopolist overestimated the price elasticity of demand by a factor $1/m$.

Substituting (5) into (1) we find the quantity demanded in the second best situation

$$q=g\left(\frac{n}{m+n}\right)^n\left(\frac{v-bx}{v}\right)^n=\left(\frac{n}{m+n}\right)^n q'$$

where q' is the quantity which would be demanded under marginal cost pricing and deficit finance. This is the equiproportional reduction formula.

Equation (5) is a two-part tariff, but not of the simple marginal cost plus tax variety. The divergence of prices from marginal cost is accomplished by a tax-cum-subsidy that minimizes the welfare cost of the distortion. Therefore one may say that the tariff really has three parts. The hidden third part is the discriminatory one, and the source of the tariff's efficiency. This third part is also the one which will allow us to distinguish the discriminatory tariff $t+rx$ from the competitive price $c+bx$. We smoke it out by writing (5) as

$$f(x)=bx+\frac{m}{m+n}v-\frac{m}{m+n}bx$$
$$=\text{Marginal Cost}+\text{Tonnage Tax}$$
$$-\text{Mileage Subsidy}$$

The revenue raised by the tonnage tax is applied to pay a share of the fixed cost K, after which a balance is left over and returned to shippers in the form of a mileage subsidy which reduces the conveyance rate to something less than marginal haulage cost. This is a cross subsidization of long-haul shippers by short-haul shippers. Note that, given v and b, the tonnage tax cannot be reduced unless the mileage subsidy is cut in equal proportion, and the conveyance rate is thereby increased. This relationship between the tax and the subsidy is used in Section II to identify tariffs. Since the factors of v and bx in (5) add up to unity, the effect of the tax-cum-subsidy is that a marginal shipper, who is a miles away from his market, where $v=ba$, and who has an infinitely elastic demand for transport, will pay as much in tax as he receives in subsidy, and is thus served at marginal cost.

Figure 3 will assist a verbal explanation of railway rates. The value of service is v dollars per ton, assumed constant at every point of a line. Marginal transport cost per ton bx increases with distance as shown by the ray $0Q$. If to this were added a per ton mile toll —such as the fuel tax—one would obtain a tariff such as that indicated by the line $0T$.[2] This would restrict traffic to movements over distances equal to or less than $0e$. Additional traffic could be carried to points between e and a if price were lowered to a level more nearly equal to marginal cost. Moreover, the line $0T$ represents inefficient pricing in case there were fixed quantities to be shipped from every point on the road. In that case, one could obtain the same amount of toll revenues by a tonnage tax, raising the price of short hauls, and reducing the price of long hauls, as by a tariff such as the broken line through point R, parallel to $0Q$.

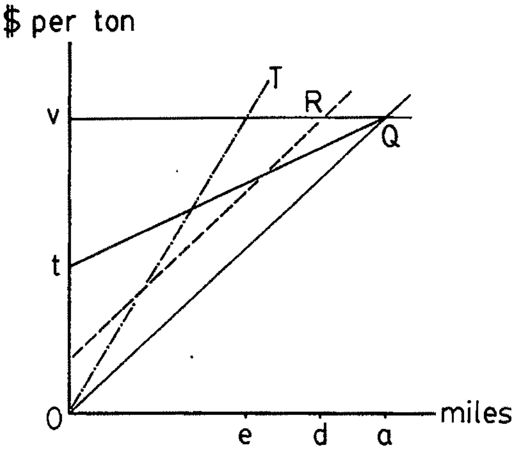


FIGURE 3. ALTERNATIVE TARIFFS

Traffic is now extended to point d, and on this basis Walters condemned the fuel tax in favor of a tonnage tax. However, this still is inefficient. By raising the tonnage tax again —this time not to pay for the road but to subsidize its more extended use—one can set the tariff illustrated by tQ. That is what a monopolist would do—as per equation (5)—especially if quantities depend on price, because he would then reduce the price of long hauls, the demand for which is elastic, and raise the price of short hauls, where demand is inelastic.

The profit-maximizing tariff (with $m=1$) was found by Launhardt in 1890, together with the inverse elasticity rule (p. 45). The same result was obtained in 1840 by Ellet for the special case of linear demand curves ($n=1$). In this last case, the conveyance rate is reduced to one-half of marginal cost per ton mile. This is like a spatial monopolist's absorption of half the freight, which Hans Singer rediscovered in 1937.

Launhardt was aware of a sustainability problem. He noted that terminal charges allow railway rates to be undercut by higher cost operators (p. 46). To see this, think of the line $0T$ in Figure 3 as a truck's constant cost, including the inefficient fuel tax it pays. Over short distances, trucks can charge less per ton than railways, even if they have higher average costs per ton mile. They can do this because of the inefficient structure of road-user charges.

[2] This line illustrates what is known in the railway literature as a *pro-rata* freight rate.

Launhardt did not recommend his profit-maximizing formula for practical application, since for most distances it would have raised rates far above their then current levels (p. 67). But why would rates have been so low? Normally, there is a constraint on firms, subjecting them to a perfect competition which makes them act as if the demand for the product were infinitely elastic. In this case of firms on the brink of deficit finance, a gentler constraint makes the firm act as if it only overestimated the elasticity of demand. This could be accomplished by indirect "competition of markets," by means of railways, to which these have to adjust.[3] The Interstate Commerce Commission (ICC) disparaged the concept, calling it a "euphemism for railway policy,"[4] without saying what the policy was. It was the policy of a member in a cartel which assigns customers to firms. These customers compete with those of other transport firms. A railway will then find it profitable to grant rebates which assist its customers to increase their share of distant commodity markets. Since a railway's customers compete not only with those of other railways, but also among themselves, rebates are governed by a sort of "most favored nation" clause by which the firm extends to all its customers the rate reduction granted any one of them. This clause puts downward pressure on all rates and is implied in the managers' claims—repeated to every investigating committee of Congress or Parliament—that they would not resort to local and personal discrimination, except to meet *direct* competition between common points. These discriminations were a manifestation of the unsustainability of railway rates, and therefore one of the main concerns of railway cartels and the ICC.

The idea of *indirect* competition putting a constraint on the firm is of course somewhat vague, and could therefore be expressed in many ways. Mr. Thomson, manager of the Pennsylvania Railroad, explained it in a letter to his shareholders dated November 5, 1861:

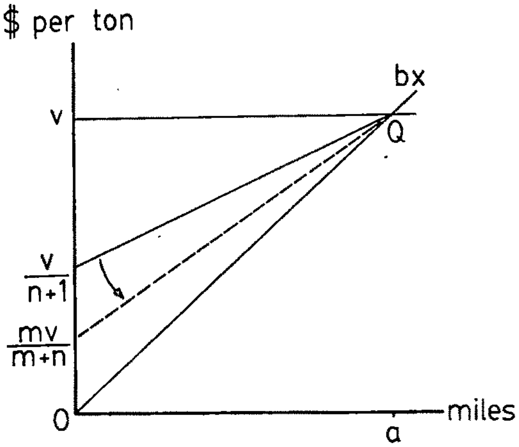


FIGURE 4. MODERATION OF MONOPOLY POWER

> For all through traffic, competition, in times of peace, limits our rates at a sufficiently low figure. But on our local traffic this competition to a large extent ceases, and the rates are fixed with reference to the market value of the article at the point it is to be shipped to, and the cost of producing and delivering it at the railroad station (if the transportation of it yields a net profit to the carrier), a margin being allowed sufficient to induce the operator or merchant to enter into and continue the business.

A monopoly railway's reaction to competition of markets is to reduce the tonnage tax and withdraw part of the mileage subsidy, as if there had been an increase in n. It would cut the terminal charge t, and raise the conveyance rate r, so the tariff line is rotated around the point Q, where marginal cost equals value of service. This would reduce the ratio t/r, and recommend it as a measure of exercised monopoly power. Across the board cuts are ruled out, because the marginal shipper is already getting service at marginal cost. (See Figure 4.)

A striking example of a tariff's rotation around Q was given by the first fixing of local rates by the South Carolina Railroad Commission in 1883. They did it in the

[3] See Royal Commission on Railways, Question 14, 795; John Clark, p. 51; and Trevor Heaver and James Nelson, pp. 157ff.
[4] See ICC 1911.

following way:

> 1st. The rates for short distances have been generally reduced below what they were before.
> 2d. The rates for long distances are about the same as before.
> 3d. In the case of the Charlotte, Columbia, and Augusta Railroad, short distance rates have been increased and long distance rates reduced. [p. 8]

In the last case, the rates—which had previously been set by the Courts—were revised because they had been strictly proportional to distance, like the cost plus fuel tax condemned by Ellet and Walters.

II. A Test of the Cross-Subsidization Hypothesis

Competitive and discriminatory transport prices are deceptively similar, since both may consist of two parts. In this section I test the theory that railroads had a cross-subsidizing discriminating monopoly. The test will be carried out on the basis of price information alone. Ordinarily, this would be rather foolish, but not in this case, because two-part prices signal a little more information than uniform prices, just enough more to permit this test.

Consider first the competitive case. Competition would make both parts of the price tend towards their costs. The firms' estimates of these costs suffer from random errors, so there is no reason to suppose that errors in terminal charges will be correlated with those in conveyance rates. A firm overestimating one of them may over- or underestimate the other, or even get it right. Therefore, if the two parts of the price were plotted on a scatter diagram, one would get a cloud around their means.

However, transport is a spatial industry, and some firms operate in high-cost areas while others serve low-cost areas. Because of the common effect of factor cost differences on c and b, one would then expect a significantly positive correlation between the two parts of a competitive transport price.

A nondiscriminating monopoly taxes all its customers at the same rate. If its terminal cost c were zero, its rates would be like the line $0T$ in Figure 3. If it had a positive terminal cost, it would mark it up by the same percentage as the line-haul cost. Variance of monopoly power among firms would then again produce a positive correlation of terminal charges with conveyance rates.

Equation (5) however implies a negative correlation between t and r, given v and b. Among profit-maximizing firms, the seesawing of the tariff parts is brought about by variance of n in the sample. Among constrained firms, variance of b, g, K, n, and v causes a variance of profits which the constraint equalizes and grinds into a variance of m. In either case, the hypothesis to be tested is that the coefficient of partial correlation between t and r is negative, holding v and b constant.

The two tariff parts in equation (5) can also be written as t and $b(v-t)/v$. Their negative correlation would be perfect, if both b and v were constant. But they are not, and if their variances spoke louder than the cross subsidization, we will find positive correlations of t and r. This can be verified by looking at tariffs applied by one and the same railway to different commodities.

Considering the tariffs of different firms for a given commodity or class, one would have to control for cost differences among firms. This I shall not do—except by grouping of data—and therefore the correlation of t with r will test both the hypothesis of cross subsidization, and the additional hypothesis that this is the strongest and most pervasive element in the determination of railway rates.

The data are local station-to-station rates and mileage scales which were step functions of distance, published by state railroad commissions. In the jargon of the trade, the local rate is that charged for the transportation of goods between points served by one and the same company, whereas through rates are applied to the interchange of traffic between companies. The geography of railway networks is such that through rates are more likely to have been made competitively than local rates, because the interchange is made at nodes where there is a choice of routes and companies. But the local traffic between nodes is also competitive. While there is no such thing as a pure monopoly tariff, the local rate is the closest substitute for it. The characterization of through rates as competitive, and local traffic as monopolized, is borne out by Thomson's letter, quoted above. Therefore I use local rates to test the model, and the model will fail if they were set with "too many" competitive considerations.

The prices shown in the tariff schedules were regressed on the distances between stations, or on the higher of the two distances in each mileage block, to estimate the terminal charge t and the conveyance rate r of each company in the sample. I obtained local rate schedules of 144 companies, operating in ten states over forty years. Of these schedules, I used mostly Class I rates, because although all companies had different numbers of classes and many commodity rates, they were not easily comparable among firms, unless they applied a common classification. But Class I is comparable, in the sense of being the highest rate a firm thinks it can charge for any commodity. Class I rates may nowadays be regarded as obsolete, but in the pre-ICC era, 25 percent of westbound tonnage of six major trunk line railroads was rated in Class I. (See New York Railroad Commissioners, p. 115.)

The tariffs were published at different times. Some state railroad commissions published the tariffs they found at the time they commenced their work, others published the rates they attempted to enforce. To the extent that these local rates applied to intrastate traffic, they were beyond the reach of the ICC. The state commission reports indicate whether or not a state had the power to set rates, and on this basis the tariffs were classified as regulated or unregulated.

A total of 2,189 unregulated prices were fitted to 146 tariff equations, and another 1,095 regulated prices were used to estimate 62 tariffs. A surprisingly large number of rates fit very well the straight line distance functions. The simplifying assumptions built into the model to produce these straight lines seem to have done little violence to facts. Also, all but 2 of 208 tariffs had significantly positive terminal charges.

Next, analyses of covariance were made to test the hypothesis that the individual tariff lines were all drawn from the same population. If they were, the seesawing of the tariff parts would have been produced by errors in the equation, and not by systematic cross subsidization. In the cases where this hypothesis could be rejected, I proceeded to correlate $\hat{t}$ with $\hat{r}$ to test the cross-subsidization hypothesis. These correlations are of $\hat{t}$s and $\hat{r}$s applied during a given year in a given state, or by a group of geographically close railways. I had data to form twenty-four such groups.

The test results are shown in Table 1. The hypothesis that all tariffs in a state are identical can be rejected, even when comparing local rates applied by one and the same company over its different divisions. The cases in which it can be accepted are those of the largest railways in states whose regulating commissions had the power to set rates, and set them uniformly for a number of companies at a time. The level of confidence with which the hypothesis of zero correlation between $\hat{t}$ and $\hat{r}$ can be rejected is generally quite acceptable, considering that we are testing a hypothesis which is stronger than necessary. The correlations always have the expected negative sign.

A more direct but statistically less satisfactory test is made by showing that when competitive price reductions are made, the percentage reduction on short hauls is greater than for long hauls, or that terminal charges are mushy and tend to disappear in the railways' attempt to obtain competitive business.

In this connection it is interesting to note that the rates set by the trunk line cartel for points between Pittsburgh and the Mississippi River were made as percentages of the New York-Chicago rate. The percentage points were initially set strictly proportional to distance (implying $t=0$), but subsequently altered to reflect an assumed terminal charge in the New York-Chicago rate. The proportionality to distance was, however, retained for points west of Chicago, where competition was most intense. This system of rate making, best described by William Ripley, was sustained from 1874 to 1931, until the ICC condemned it. Paul MacAvoy has shown that the cartel's pricing policies invited entry. Their rates must therefore have been above marginal cost, and can be represented by the line $0T$ in

TABLE 1—TESTS OF CROSS SUBSIDIZATION IN LOCAL RATES OF RAILWAYS

		F-test of Deviations between Individual Tariff Equations			Correlation of Terminal Charges with Conveyance Rates	
Year	Group	Degrees of Freedom	F	Confidence Level	$R_{\hat{t},\hat{r}}$	Confidence Level
Unregulated Railways						
1867	Ohio–Class I					
	Group 1[a]	8/34	1.024	–[d]	–	–
	Group 2[b]	8/41	8.534	99.9	−0.845	99.3
	Group 3[c]	4/38	3.902	99.0	−0.643	75.3
1874	Illinois–Class D	12/188	3.706	99.9	−0.664	98.6
	Coal	15/250	4.856	99.9	−0.601	98.4
1886	Mississippi–Class I	15/204	11.084	99.9	−0.670	99.0
1891	North Carolina–Class I	8/160	32.213	99.9	−0.926	99.9
1898	Michigan–Class I	25/206	3.970	99.9	−0.594	99.6
1900	Arkansas–Class I	5/106	22.18	99.9	−0.862[e]	85.0
1906	Indiana–Class I:	36/670	18.97	99.9	−0.693	99.9
	C.C.C. & St. L. R.R.	6/114	2.101	94.0	−0.510	75.0
	P.C.C. & St. L. R.R.	5/73	22.45	99.9	−0.965	99.8
	C.I. & L. R.R.	2/51	10.43	99.9	−0.754	54.0
	Vandalia R.R.	2/69	98.21	99.9	−0.834	60.2
	Another 18 firms	17/363	14.91	99.9	−0.409	90.6
Regulated Railways						
1897	Florida–Class I	6/100	17.429	99.9	−0.857	98.5
1905	Alabama–Class I					
	Large RRs.	2/81	1.794	–	–	–
	Medium	6/136	6.47	99.9	−0.253	41.6
	Small	10/99	7.115	99.9	−0.475	85.2
1907	Mississippi–Class I					
	Large	5/266	4.976	99.9	−0.669	84.4
	Medium	6/156	6.70	99.9	−0.730	93.4
	Small	5/23	16.60	99.9	−0.240	35.2
1907	So. Carolina–Class I					
	Large	2/69	2.649	91.0	–	–
	Small	11/41	6.236	99.9	−0.944	99.9

[a] Group of unprofitable railways.

[b] Dividend paying companies, and railways leased to dividend paying ones.

[c] Four of the five in the group were large dividend paying railways.

[d] The tariffs in this group were parallel lines.

[e] Calculated after eliminating one tariff which was not significantly different from another. Eliminating the other one instead, $R_{\hat{t},\hat{r}}$ is −0.996.

Figure 3. Thus competition among members of an unstable cartel brought about a system of rate making which Pigou, Ramsey, and Hotelling would have condemned, had they been commissioners and Ellet, Launhardt, and Walters their staff members.

III. Testing the Presence of a Constraint on Profit

In the tariff equation $f(x)=mv/(m+n/+nbx/(m+n)=t+rx$, the ratio of terminal charge to the conveyance rate, t/r, equals mv/bn, or am/n, where a is the maximum economically possible length of haul—determined by demand and cost conditions beyond the firm's control—m is an index of the constraint on the firm's profit which varies from zero (no monopoly) to unity (no constraint), and n is the elasticity of demand with respect to the proportion of the value of service conceded by the carrier to the shipper. Thus one may say that m measures the firm's power to tax, and n measures the revenue that the tax will not raise, so m/n is the railway's monopoly power and t/r is a good index of it.

TABLE 2—LOCAL RATES IN OHIO, CLASS I, 1867

Group of Railways	$\hat{t}$	$\hat{r}$	R^2	$\hat{t}/\hat{r}$
Unprofitable	15.097 (0.793)	0.2452 (0.0072)	0.9586	62
Dividend Paying Railways, and Unprofitable Railways Controlled by Profitable Railways	9.778 (0.466)	0.2303 (0.0045)	0.9787	42
Mostly Dividend Paying	8.375 (0.488)	0.1955 (0.0049)	0.9720	43

Source: Calculated from Ohio, Commissioner of Railroads and Telegraphs.
Note: Standard errors of coefficients are shown in parentheses. Rates are in cents per 100 pounds.

TABLE 3—CONVEYANCE RATES AND TERMINAL CHARGES ON LOCAL FREIGHT BEFORE REGULATION (BEFORE 1887), AND DURING PERIODS OF STRONG (1888–95) AND WEAK (1896–1902) REGULATION: RATIO OF TARIFF PARTS $\hat{t}/\hat{r}$

Years	Railway and Point of Origin			
		Michigan Central		
	Chicago	Kalamazoo	Detroit	
1884	164	253	162	
1887–95	153	58	81	
1896–1902	216	189	202	
		Chicago and Grand Trunk		
	Chicago	Port Huron	Flint	Charlotte
1882	99	162	95	132
1887	102	251	162	245
1890	128	92	44	32
1895	128	114	76	44
1900	172	213	120	120
		Lake Shore & Michigan Southern		
	Chicago		Toledo	
1886	184		89	
1887–95	150		70	
1896–1902	196		120	

Source: Calculated from rate tables in ICC, 1903.

Among profit maximizers, a low n makes a high t/r and a high profit in equation (6). This correlation between t/r and profits is spoiled by the constraint, which only allows high taxes (high t/r) to poor railroads. Table 2 provides a rough illustration of this.

Consider now the adjustment of rates when there is an untaxed sector as well as a taxed one. In the case of railways, we may regard long-distance through traffic as lightly taxed, and short-distance local traffic as heavily taxed. If there was a constraint on profits, any gains reaped by a cartel's pooling of competitive through traffic must have been at least partially offset by relief to shippers in noncompetitive local markets. This is one basis for railway officers' and some shippers' representations to Congress in favor of pooling (see Gabriel Kolko, p. 77). To the officers it had the advantage of changing the competition's character, making it less direct, as by potential entry and rivalry of markets. Their arguments implied a negative correlation of through rates set by cartels with the local rates of cartel members. Some information on this point is shown in Table 3. The Michigan Central, Grand Trunk, and Lake Shore were members of a cartel allegedly strengthened by the Interstate Commerce Act, especially Section 4, the long- and short-haul clause. This clause was increasingly disregarded as time wore on. It was challenged in the courts and rendered ineffective by the Supreme Court's decision on the Alabama Midland case in 1897. Note that this clause did not order a reduction of terminal charges and an increase in conveyance rates, and could have been observed by revising only the latter (for example, by "blanketing" or setting $r=0$). Nevertheless, as shown in Table 3, three cartel members charged their local traffic with a lower t/r during the period of strong regulation around 1890 than either before or after it. Also, this index of exercised monopoly power was often higher in 1900, when traffic pooling by a cartel was illegal, than in the 1880's, when it was not.

Now suppose a railway line was providing only local service, charging the tapered rates of equation (5). Suppose also that the demand intercept g is constant at all points on the line, and the surplus revenue over direct cost is as in equation (6). If this line was turned into a portion of a longer through route, the balance of long- and short-haul traffic would be changed, as would the distribution of g over the line. It was noted in Section I that the distribution of g has no effect on the tariff on a profit-maximizing railway. The company will join the through route if the revenue from through traffic exceeds whatever additional expenditures the new service requires. But if there is an effective profit constraint, the level of the local tariff (the size of m) will depend on the distribution of traffic between long and short hauls. A railway asked to make many low-priced long hauls and a few high-priced short hauls will have a higher m than a company specializing in short hauls.

This hypothesis of constrained adjustment of local rates to the relative size of the through business is tested with data concerning the case of Ohio, from July 1, 1881 to June 30, 1888. Let X be ton miles of long-distance through traffic, and s the corresponding surplus per ton mile over direct cost. The surplus revenue from through traffic, sX, is proportional to output because the cartel which regulated the traffic through Ohio set rates on a *pro-rata* basis, and the revenue from joint traffic was prorated among the companies on a mileage basis. The total surplus over direct cost of a cartel member was sX plus the net revenue from local traffic in equation (6):

$$S=sX+\frac{mZ}{(m+n)^{n+1}}\qquad Z=\frac{n^n g v^2}{(n+2)b}$$

If indirect competition creates a constraint on S, then $S=K$ and

$$dk=dS=s\,dX+Z\frac{(m+n-mn)}{(m+n)^{n+1}}dm=k\,dX$$

If increased through traffic can be handled without any increase in common costs, then $k=0$. The percentage of variable cost would then decline as traffic increases, whereas the actual experience of railways was that K rose *pari passu* with output. Their increasing returns to scale arose not out of a simple constancy of K, but have the subtler reasons explained by David Haddock in terms of increasing returns to train size and the cost of scheduling less frequent trains. But if $k>0$, the change in a railway's power to tax local traffic allowed by a change in its lightly taxed through traffic is

$$\frac{dm}{dX}=\frac{(k-s)(m+n)^{n+1}}{Z(m+n-mn)}$$

The sign of this derivative is the same as the sign of $(k-s)$. Should competition drive s towards zero, an increase in through traffic will mean a heavier tax on the local. If a cartel raised s above the level of k, m may be reduced by the profit constraint and the through traffic will subsidize the local. If there were no profit constraint, railways would set the tax on local traffic at its profit-maximizing level and leave it there. Therefore, a significantly positive correlation between the tax on local traffic and ton miles of through traffic would show the presence of a profit constraint, and a subsidy of through by local traffic.

TABLE 4—RELATIVE LEVELS OF TRAFFIC AND RATES IN OHIO, 1882–88

Year Ending June 30	c_0	c_1	c_2	c_3	$\bar{R}^2$
1882	1.933	−0.002288 (−1.136)	0.2879 (5.168)	−0.0945 (−2.099)	0.7192
1883	1.797	−0.003367 (−2.559)	0.1322 (3.860)	−0.0015 (−0.022)	0.5512
1884	2.207	−0.004644 (−2.820)	0.0355 (2.042)	−0.0238 (−0.560)	0.3902
1885	2.462	−0.004776 (−2.811)	0.0876 (2.077)	−0.1346 (−2.040)	0.3680
1886	2.510	−0.003514 (−1.928)	0.1284 (2.443)	−0.2047 (−2.921)	0.3574
1887	1.937	−0.006700 (−3.861)	0.2802 (6.840)	−0.3057 (−5.212)	0.7270
1888	2.053	−0.005624 (−2.952)	0.1351 (2.696)	−0.0085 (−1.701)	0.4242

Note: *t*-ratios are shown in parentheses.

Let Y be the ratio of the average revenue per local ton mile relative to the average revenue per ton mile of through freight. Since the denominator was close and proportional to marginal cost, Y is a rough measure of the tax on local traffic and a proxy for m. Since Y corresponds to an average tax rate, yet taxes are graduated in inverse proportion to length of haul, we must expect Y to be negatively correlated with X_1, the average length of the local haul. Let X_2 be the ratio of through ton miles to local ton miles. This corresponds to the X in the constraint and is measured in ratio form so we may use cross-section data. Finally, we also add X_3, the gross revenue from freight per dollar of passenger revenues, to catch another possibly untaxed sector, and write the regression equation

$$Y = c_0 + c_1 X_1 + c_2 X_2 + c_3 X_3 + u$$

If c_2 or c_3 is significantly different from zero, that will be evidence of cross subsidization produced by a profit constraint. The numerical values of the coefficients do not mean as much as their deviation from zero, since they are functions of m, the constraint would give each firm its own m, and different firms are represented in each cross-section.[5]

The estimates of the coefficients in the regression equation may suffer from an identification problem, since substitution in consumption causes positive correlation of Y with X_2. But in this case it is difficult to argue that the traffic of Western States with the Eastern Seaboard through Ohio was a close and statistically significant substitute for local traffic in the state. Also, the coefficient of X_2 will be biased downwards, because Y equals X_2 times the ratio of gross revenues from local and through traffic.

The estimates are shown in Table 4. The coefficient of the average length of haul has the expected sign. The signs of the other coefficients indicate that the local freight subsidized every other line of the business. In 1886–87 this was so bad that Cincinnati, Hamilton, & Dayton; Cincinnati, Richmond, and Chicago; Cincinnati, Hamilton, & Indianapolis, and the Dayton and Michigan Railroads withdrew their participation in through traffic. (See Cincinnati, Hamilton, & Dayton Railroad Co., p. 11)

[5] Few companies and railroad commissions reported through and local traffic separately. The Ohio commissioner published tonnages, ton-miles, and revenues of through and local traffic from 1885 to 1888. Data for 1882, 1883, and 1884 were obtained solving appropriate systems of equations with less explicit information from the same source. Ohio had eighty-nine railways in 1882, seventy-one in 1888. Forty-two of these made useable reports at one time or another; I used at most twenty-seven in any one year. Not all of these were independent companies. In 1886, for instance, nine were leased to or owned by some of the other eighteen companies.

TABLE 5—RELATIVE LEVELS OF TRAFFIC AND RATES: SELECTED FIRMS

Period	Railway	c_0	c_1	c_2	$\bar{R}^2$
1869–88	Cleveland, Columbus, Cincinnati, & Indianapolis	1.7077	−0.00432 (−0.399)	0.2026 (2.756)	0.3151
1885–1900	New York Central	1.3220	−0.00634 (−0.537)	0.2317 (3.116)	0.8214
1881–95	Lake Erie & Western	2.0374	−0.00701 (−1.694)	0.2677 (1.463)	0.3434
1885–98	Cleveland, Akron, and Columbus	3.4922	−0.03471 (−3.671)	0.1085 (0.507)	0.6243
1885–1900	Erie	0.9587	0.00522 (1.460)	−0.1109 (−1.41)	0.0973
1885–1900	Grand Trunk	2.8853	−0.00836 (−1.54)	−0.0511 (−0.06)	0.0252

Source: Calculated from data in annual reports of the companies and of railroad commissions of Michigan, New York, and Ohio.
Note: t-ratios are shown in parentheses.

A time-series analysis can produce different but consistent results. Following individual companies through time, we find the same positive relationship between Y and X_2 holding for profitable railways like the New York Central and the Cleveland, Columbus, Cincinnati, and Indianapolis. This relationship persists, but is much weaker in the case of moderately successful companies like the Lake Erie and Western and the Cleveland, Akron, and Columbus, which emerged from financial difficulties by reorganization in the 1880's. But the Erie and Grand Trunk, both notorious for their financial troubles, show the zero correlation expected of poor railroads allowed to maximize profits. This is shown in Table 5, where the passengers have been omitted from the equation, and c_1 is not significant because the average length of haul by any railway was virtually constant over time.

REFERENCES

W. J. Baumol and D. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

John M. Clark, *Standards of Reasonableness in Local Freight Discriminations*, New York 1910.

R. H. Coase, "The Marginal Cost Controversy," *Economica*, Aug. 1946, *13*, 169–82.

R. S. Damus, "A Two-Part Ramsey-Optimum Railroad Tariff," unpublished doctoral dissertation, Univ. Chicago 1979.

C. Ellet, Jr., "A Popular Exposition of the Incorrectness of the Tariffs on Tolls in Use on the Public Improvements of the United States," *J. Franklin Inst.*, 1840, *29*, 225–32.

D. D. Haddock, "Regulation of Railroads by Commission," unpublished paper, Univ. Chicago, Feb. 1978.

Authur T. Hadley, *Economics: An Account of the Relations Between Private Property and Public Welfare*, New York 1896.

G. R. Hawke, "Pricing Policy of Railways in England and Wales Before 1881," in M. C. Reed, ed., *Railways in the Victorian Economy*, New York 1968, 76–110.

Trevor D. Heaver and James C. Nelson, *Railway Pricing Under Commercial Freedom: The Canadian Experience*, Vancouver 1977.

H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, *6*, 242–69.

Gabriel Kolko, *Railroads and Regulation, 1877–1916*, New York 1970.

Wilhelm Launhardt, *Theorie der Tarifbildung der Eisenbahnen*, Berlin 1890.

Paul W. MacAvoy, *The Economic Effects of Regulation: The Trunk-Line Railroad Cartels and the Interstate Commerce Commission Before 1900*, Cambridge, Mass. 1965.

S. Peltzman, "Towards a More General Theory of Regulation," *J. Law Econ.*, Aug. 1976, *19*, 211–40.

A. C. Pigou, *Economics of Welfare*, London 1919.

R. A. Posner, "Taxation by Regulation," *Bell J. Econ.*, Spring 1971, *2*, 22–50.

F. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47–61.

W. Z. Ripley, "The Trunk Line Rate System: A Distance Tariff," *Quart. J. Econ.*, Feb. 1906, *20*, 183–210.

H. W. Singer, "A Note on Spatial Price Discrimination," *Rev. Econ. Stud.*, Oct. 1937, *5*, 75–77.

G. S. Stigler and C. Friedland, "What Can Regulators Regulate? The Case of Electricity," *J. Law Econ.*, Oct. 1962, *5*, 1–17.

Alan A. Walters, *The Economics of Road User Charges*, Baltimore 1968, ch. 5.

Cincinnati, Hamilton, & Dayton Railroad Co., *Annual Report of the Directors for the Year Ended March 31, 1887.*

Interstate Commerce Commission, *Railways in the United States in 1902, Part II, A Forty-Year Review of Changes in Freight Rates,* Washington 1903.

______, *Railroad Commission of Nevada vs. Southern Pacific Company et al.*, 21 ICC 367 (1911).

______ *vs. Alabama Midland Railway*, 168 U.S. 144 (1897).

New York Railroad Commissioners, *Report for 1885*, Vol. I.

Ohio, Commissioner of Railroads and Telegraphs, *Annual Report for the Year 1867,* Columbus 1868.

Pennsylvania Railroad Company, Committee of Shareholders Appointed February 4, 1860, to Investigate the Condition and Policy of the Pennsylvania Railroad, *Minority Report*, Philadelphia 1862.

Railroad Commissioners for the State of South Carolina, *Fifth Annual Report*, Columbia 1883.

Royal Commission on Railways, *Evidence Taken Before the Commission March, 1865 to May, 1867,* London 1867.

Workmen's Compensation and Occupational Safety under Imperfect Information

By SAMUEL A. REA, JR.*

A wide range of government policies indicate a lack of appreciation by policymakers that market mechanisms might provide optimal insurance and safety. This distrust of market outcomes is particularly evident in programs such as workmen's compensation and the Occupational Safety and Health Act (OSHA). Many economists have justified mandatory insurance and safety regulation by suggesting that several types of imperfect information adversely affect the performance of insurance and safety markets, but there has been little analysis of the effects of government intervention when imperfect information is present. This paper develops a common framework for analyzing market-provided disability insurance, workmen's compensation, and occupational safety legislation in the presence of imperfect information.

There are five possible types of imperfect information which affect the market for disability insurance and occupational safety: 1) Employees may be incorrect in their estimates of occupational risk and their influence on the level of risk; 2) it is difficult for the employer to monitor the precautions taken by employees; 3) the workmen's compensation board or insurance carrier cannot monitor employers' or employees' precautions; 4) employers may not be able to identify workers who are accident-prone; 5) the insurance carrier may not be able to monitor the extent of injury. The second, third, and fifth types of misinformation are commonly called moral hazard, and the fourth type leads to adverse selection. The first two types of misinformation are considered in this paper with particular attention paid to worker underestimation of risk.

Economists since Adam Smith have suggested that consumers and workers underestimate risk.[1] This underestimation of risk is often cited as a justification for the regulation of safety and insurance. For instance, mandatory use of seat belts is seen as optimal because drivers do not correctly evaluate their effect on the expected loss. Food substances are banned because many consumers are said to ignore health risks. Mandatory insurance accompanies no-fault legislation because it is believed that people will not insure themselves. Walter Oi, Peter Gregory and Micha Gisser, Albert Nichols and Richard Zeckhauser, and Peter Diamond have argued that misperceptions justify workmen's compensation and occupational safety regulation.

In the employment context, the argument for mandatory insurance suggests that the expected utility of a worker, evaluated using the true probability of an accident, will be increased by mandatory insurance. This argument neglects an important consideration: the regulation of insurance does not alter the worker's misperception of risk. Consequently he will desire to adjust other unregulated quantities such as safety precautions in response to the regulation of insurance coverage. Responding to the worker's preferences, the firm will vary the compensation and working conditions, including the firm's safety precautions and the precautions expected of the employee. The model developed in this paper indicates that safety could fall as a result of mandatory insurance (workmen's compensation) even if there is no moral hazard. This conclusion contradicts the predictions of Oi and of

*Associate professor of economics, University of Toronto. This research was supported by the Ontario Economic Council, but the study reflects my views and not those of the Council.

[1] Adam Smith noted that "the chance of loss is by most men undervalued" (p. 107). He discusses the unusually low percentage of people who buy fire insurance (p. 108) and the underestimation of risk by young people choosing their occupations (p. 109).

Diamond (pp. 80–81). If safety falls sufficiently, workers' utility, evaluated with the true probabilities, could be lowered by workmen's compensation.

The model developed below focuses on the compensation package that will be offered to workers in an industry with homogeneous workers and firms. This compensation package includes the wage, disability benefit, a level of precautions taken by the firm, and a level of precautions expected of the worker. The optimal levels of safety and insurance, derived in the first section, will be provided in the market if there is perfect information. The model is used to examine the effect of underestimation of risk on the market outcome. Finally, the impact of government regulation of insurance and safety is analyzed.

I. Optimal Insurance and Safety with Perfect Information

Workers are not only interested in wage payments, they are also concerned with payments in the event of disability and the probability of disability. The probability of disability is reduced by precautions taken by the worker and by the firm, but these precautions are costly to both. The employee is assumed to maximize his expected utility, which is a function of the probability of disability, the utility if no disability occurs and the utility if he is disabled. Utility in each state of the world is a function of income and safety precautions. Since precautions require additional effort or concentration, they reduce utility in each state of the world. Precautions are undertaken because they increase *expected* utility. If the worker is not injured, he receives his wage w, but if he is disabled, he receives a disability benefit m. The labor supply is assumed fixed. The expected utility is

$$(1) \quad a(d,s)U(w,d)+(1-a(d,s))V(m,d)$$

where

$a(d,s)$ = probability of no disability
d = index of worker's safety precautions
s = index of employer's safety precautions
w = wage
m = disability benefit
$U(w,d)$ = utility if not disabled ($U_w<0$, $U_{ww}<0, U_d<0, U_{dd}>0$)
$V(m,d)$ = utility if disabled ($V_m<0, V_{mm}<0, V_d<0, V_{dd}>0$)
$U>V$ if $w=m$

Safety precautions can enter into the firm's profit equation in at least four ways. First, they affect the supply of labor to the firm and the wage and benefit package that is offered. Second, they impose costs that may be associated with each job slot. (Protective clothing would be the most straightforward example of this type of precaution.) Third, precautions may be associated with capital, such as a protective device on a machine. Fourth, the precautions may enter the production function directly. (For instance, a lower operating speed will reduce output but will lower the probability of an accident.) In order to simplify the analysis it is assumed that labor is the only factor in the production function. Following Diamond, precautions are assumed to affect the fixed cost of hiring an employee, but are invariant to the level of employment.

Firms will compete for labor by varying the components of the compensation package, broadly defined to include the firm's safety precautions s and the level of employee precautions required d, in addition to the wage w and the disability benefit m. It is assumed that the number of employees in the industry is predetermined. Diamond assumed that the firm or industry has to compete with a riskless industry for labor. Use of this approach produces the same optimal conditions, but complicates subsequent analysis. Under Diamond's assumptions, the effect of government regulation of insurance or safety is to alter the supply of workers to the risky industry and alter the relative prices of the goods being produced, making the distributional effects of the regulations difficult to trace. The approach taken here imposes all of the costs of a policy that regulates insurance or safety on the workers in that industry.

The firm hires workers knowing that some proportion of the workers will become disabled and will be nonproductive. Workers are assumed to have homogeneous preferences and productivity. Workers are not paid a wage if they become disabled, but the firm will have to pay disability benefits and will have other costs associated with the disabled worker's job slot. These costs include any specific investment in human capital. Entry of firms into the industry and competition for employees (which varies w, m, d, and s) will guarantee that the expected marginal product of labor equals the expected marginal cost:

$$a(d,s)Z=A(d,s)w+(1-a(d,s))m+C(s)$$

or

$$(2)\quad Z(a)=w+\left(\frac{1-a(d,s)}{a(d,s)}\right)m+\frac{C(s)}{a(d,s)}$$

where $Z(a)=$ marginal revenue product of labor ($Z'<0$), and

$C(s)=$ fixed costs per worker employed ($C'(s)>0, C''(s)>0$).

The marginal revenue product of labor increases as the risk level rises because lower levels of safety reduce the number of employees who actually produce.

The socially optimal level of disability benefits and safety can be determined by maximizing the expected utility of each worker G, with respect to m, d, and s, subject to equation (2).

(3)

$$G=a(d,s)U\left(\left(Z(a)-\frac{(1-a)}{a}m-\frac{C}{a}\right),d\right)$$
$$+(1-a(d,s))V(m,d)$$

The first-order conditions are

$$(4)\quad U_w=V_m$$

$$(5)\quad -[aU_d+(1-a)V_d]=a_d[U-V]$$
$$+\frac{a_d(m+C)U_w}{a}+aU_wZ'a_d$$

$$(6)\quad C'=\frac{a_s(U-V)}{U_w}+\frac{a_s(C+m)}{a}+aZ'a_s$$

Equations (4), (5), and (6) determine the optimal levels of m, d, and s. Equation (4) represents the condition for optimal insurance given d and s. The marginal utilities are equated between both states of the world. The optimal level of precautions taken by the individual (equation (5)) should be such that the expected marginal cost of the precautions equals the marginal reduction in the uncompensated loss, plus terms that represent a reduction in the amount of disability compensation, the declining marginal productivity of labor, and the cost to the firm of the job slot. The last two terms reflect the greater wage that could be paid if the job is made less hazardous because of worker precautions.

The third condition (equation (6)) requires that the firm take safety precautions until the marginal cost of these precautions equals the marginal reduction in uncompensated loss plus the reduction in disability benefits, the costs associated with nonproductive job slots, and the decline in marginal productivity. The wage that workers are willing to forego in return for greater safety is $(U-V)/U_w$.

There are a number of possible methods for determining the level and conditions under which disability compensation (m) is paid, such as a negligence standard of care, a market mechanism, or government determination of benefit levels (workmen's compensation). The second category includes insurance purchased from a third party or from an employer. It can easily be shown that competition will produce the socially optimal level of insurance and safety if there is perfect information. When imperfect information is introduced, the equilibrium disability benefits and safety will be altered. The effect of imperfect information and government determination of benefits and safety will be illustrated with a general model that includes imperfect information on the part of workers and firms.

II. Market Determination of Safety and Insurance under Imperfect Information

The three parties in a disability insurance market (workers, employers, and insurance carriers) each have incomplete information, which leads to well-known problems in insurance: moral hazard and adverse selection.[2] In this paper all firms and employees are assumed to be homogeneous. This assumption rules out adverse selection, but does not eliminate the firm's lack of information concerning an employee's precautions. It is also assumed that the firm self-insures its liability for disability benefits. This assumption eliminates the additional moral hazard that results when the insurance carrier cannot monitor the firm's precautions. Workers may also underestimate the probability of an accident. This underestimation could account for the growing evidence that the expected utility model is not a particularly good predictor of the consumer's response to small probabilities of large loss.[3] Others have found some evidence of compensating wage differentials for occupational risk,[4] but the differentials may not equal the value of the loss of expected utility, evaluated using the true probabilities of accidents.

The incorporation of misperception in the model presented here is based on Spence's study of product liability. Workers are assumed to maximize expected utility, which depends on the worker's and the employer's precautions, but the perceived effects of these precautions may be incorrect. The employee maximizes:

$$(7)\quad r(d,s,x)U(w,d)+\left[1-r(d,s,x)\right]V(m,d)$$

where $r(d,s,x)$ = perceived probability that he will not be disabled

$$(r_i>0, r_{ii}<0, i=d,s,x)$$

where x = exogenous variable affecting employee perception.

Given a wage rate and level of employer precautions, the worker will choose a level of precautions such that

$$(8)\quad -\left[rU_d+(1-r)V_d\right]=r_d\left[U-V\right]$$

The cost of precautions is the disutility of taking additional effort, and the benefit is the increased expected utility resulting from greater safety. The worker sets the expected marginal cost of precautions equal to the expected marginal loss reduction. The term $(U-V)$ represents the loss not compensated by m.

Workers will adjust their safety precautions as the compensation package changes. Taking total differentials of equation (8) we get

$$(9)\quad \frac{\partial d}{\partial m}=\frac{(1-r)V_{dm}-r_dV_m}{D}$$

$$(10)\quad \frac{\partial d}{\partial s}=\frac{r_s(U_d-V_d)+r_{ds}(U-V)}{D}$$

$$(11)\quad \frac{\partial d}{\partial w}=\frac{rU_{dw}+r_dU_w}{D}$$

$$(12)\quad \frac{\partial d}{\partial x}=\frac{r_{dx}(U-V)+r_x(U_d-V_d)}{D}$$

where $D=-[2r_d(U_d-V_d)+rU_{dd}+(1-r)V_{dd}+r_{dd}(U-V)]>0$ if the second-order conditions hold.

It is reasonable to assume that the marginal disutility of precautions in each state of the world increases with income, $U_{dw}<0$, $V_{dm}<0$. An increase in disability compensation lowers the worker's precautions (equation (9)). This occurs because he will desire to shift income from the disabled state to the healthy state when m increases, and because the disutility of precautions in each state increases with income. An increase in the wage rate will raise precautions unless U_{dw} is sufficiently negative (equation (11)).

[2] See Michael Spence and Richard Zeckhauser, Mark Pauly, Michael Rothschild and Joseph Stiglitz, Charles Wilson, and Steven Shavell.

[3] See Howard Kunreuther et. al. and Amos Tversky and Daniel Kahneman (1974, 1979).

[4] See Richard Thaler and Sherwin Rosen, Robert Smith, and W. Kip Viscusi.

The effect of the firm's safety precautions on the worker's precautions depends in part on whether the firm precautions are perceived as complementary by the worker ($r_{ds}>0$). It is not clear a priori whether or not this is generally the case. Some expenditures by the firm may inform workers of hazards (signs, for example) and increase risk awareness ($r_{ds}>0$). Other precautions taken by the firm, such as safety devices on a machine, are likely to be perceived as substitutes for worker precautions ($r_{ds}<0$). The other term in equation (10) represents the effect of a change in the perceived level of risk on the perceived cost of precautions. It is negative if the marginal disutility of precautions is smaller in the state of the world in which disability occurs. In the restricted case where $V_{dm}=U_{dw}=0$ and $U_d=V_d$, the sign of r_{ds} determines the sign of $\partial d/\partial s$.

Firms are assumed to compete with each other for labor by offering alternative compensation packages. Firms can alter the wage, the disability benefit, and the precautions. Variations in these parameters will cause workers to vary their levels of care. A basic assumption of the model is that the firm is able to monitor the reactions of workers in the aggregate to changes in the compensation package, but is unable to monitor the precautions of specific workers. If the firm could monitor individual employee precautions, it would require a given level of precaution. The cost to the employee of required precautions would be an element in the total compensation package.

The firm sets the compensation package knowing the reaction function of the workers ($d(w,m,s)$). The firm is assumed to be self-insured against liability for disability payments. For any given level of employment, the employer offers a wage, safety, and disability benefit package that maximizes the worker's perceived expected utility,

$$(13)\quad r(d,s)U(w,d)+(1-r(d,s))V(m,d)$$

subject to the constraint that the marginal revenue product equals the marginal cost of a worker,

$$(14)\quad Z-w-\frac{(1-a)m}{a}-\frac{C(s)}{a}=0$$

where $d=d(w,m,s)$ and $a=(d,s)$. It is shown in Appendix A that by differentiating with respect to w, m, and s, and taking advantage of equation (8), we get the first-order conditions (together with equation 14):

$$(15)\quad \frac{rU_w}{(1-r)V_m}=\frac{1-\left(\frac{\partial d}{\partial w}\right)a_d\left(\frac{(m+C)}{a^2}+Z'\right)}{\left(\frac{1-a}{a}\right)-\left(\frac{\partial d}{\partial w}\right)a_d\left(\frac{(m+C)}{a^2}+Z'\right)}$$

$$(16)\quad r_s(U-V)+\left[\frac{rU_w}{1-\left(\frac{\partial d}{\partial s}\right)a_d\left(\frac{(m+C)}{a^2}+Z'\right)}\right]\times\left[-\frac{C'}{a}+\left(\frac{\partial d}{\partial s}\right)a_d\left(\frac{(m+C)}{a^2}+Z'\right)+a_s\left(\frac{(m+C)}{a^2}+Z'\right)\right]=0$$

It is obvious that the market levels of insurance coverage and safety are likely to differ from the socially optimal conditions because of the imperfect information. Because of the complexity of the model, the effects of the imperfect information are examined in four special cases: 1) market determination of insurance coverage with workers' precautions predetermined; 2) workmen's compensation (government determines m) with worker precautions endogenous; 3) workmen's compensation with worker precautions predetermined; 4) government regulation of firm precautions with insurance levels determined by the market.

A. *Market Determination of Insurance Coverage: d Constant*

If the workers have no control over the occupational risk, one source of imperfect information is eliminated. The firm does not have to take account of changes in workers' precautions when it alters the wage rate and disability benefits. Equations (15) and (16) reduce to

$$(17)\qquad \frac{U_w}{V_m}=\frac{a(1-r)}{(1-a)r}$$

$$(18)\quad \frac{ar_s(U-V)}{rU_w}=\left[C'-\frac{a_s(m+C)}{a}-a_sZ'\right]$$

Equation (17) indicates that the market level of disability insurance will be optimal only if $r=a$. If the worker is overly optimistic, $a<r$ and $U_w<V_m$, necessarily implying suboptimal insurance only if a is at the optimal level. Equation (18) indicates the safety investment criterion for the firm. Safety is increased until the marginal cost (right-hand side) equals the marginal benefit. The marginal benefit is the lower wage that can be paid if a safer job is offered. The marginal benefit differs from the marginal social benefit (equation (6)) because the worker does not correctly evaluate risk.

What are the effects of changes in worker perceptions on the market levels of m, s, and w? Consider an increase in x that raises r with $r_{sx}=0$. It is shown in Appendix A that such an increase in x will reduce the insurance coverage ($\partial m/\partial x<0$) and may increase the wage rate. This seems reasonable because workers will place a lower value on disability benefits. A surprising result is that safety may *increase*. Worker underestimation of risk leads to a substitution of wages for benefits because implicit insurance rates seem unfavorable. The lower benefits reduce the cost of accidents for the firm, but they may increase the wages workers are willing to give up in order to pay for firm precautions. The increase in safety can occur if $(1-r)/(1-a)$ is sufficiently below $(a/r)\,a_s/r_s$. If worker perceptions are initially correct, a small decrease in the estimate of risk will definitely lower insurance coverage, increase the wage rate, and reduce safety.

The results above are altered if $r_{sx}<0$. An exogenous change in perceptions that reduces r_s with r constant decreases safety as expected, but insurance coverage may increase. Insurance coverage will definitely *increase* in response to a fall in r_s when perceptions are initially correct. Firms reduce safety and increase wage rates as r_s falls, but workers will want more insurance because the perceived *level* of risk is initially unchanged. The combined effect of an exogenous increase in r and decrease in r_s is indeterminate.

The model indicates that underestimation of risk is not necessarily a justification for mandatory disability benefits. If workers underestimate the marginal impact of firm precautions, there may be too much insurance. On the other hand, if they underestimate the *level* of risk, there will be too little insurance and possibly too much safety. One must know more about the pattern of underestimation before intervention can be justified.

The ambiguity of the effect of misperception on market-determined benefits did not occur in Diamond's model because he assumed that safety was constant when he analyzed the insurance decision, and that insurance benefits were constant when he analyzed the safety decision. If s is fixed, a reduced estimate of risk lowers the insurance coverage. If m is fixed, a reduced estimate of risk lowers firm precautions. Once both insurance and safety are allowed to vary simultaneously, the results can be ambiguous and sometimes counterintuitive.

B. *Workmen's Compensation: m Legislated*

Workmen's compensation is a program that imposes strict liability on the employer for occupational disabilities, with compensation levels set by statute or regulation. In addition, workmen's compensation often includes insurance for the employer, but this characteristic does not have an important bearing on the misperception issue and is not incorporated in the model. In this section the effect of changes in the workers' estimates of risk are considered when m is held constant, and the response of safety to the level of mandatory benefits is analyzed. The mathematical derivations appear in Appendix B.

1. *Workers' Precautions Endogenous*

If the government sets benefits at a level that exceeds market coverage, equations (14) and (16) become the only first-order conditions. The zero-profit constraint requires that an increase in safety precautions by the firm be offset by a decrease in the wage rate as long as m is held constant. If the second-order conditions hold, it follows that a reduction in the workers' estimate of risk ($r_x > 0, r_{sx} \leqslant 0$) without a first round effect on d, will increase the wage rate and decrease the firm's expenditures on safety. (This assumes that $d_x = 0$, $d_{wx} = 0$ and $d_{xs} = 0$. Equations (10), (11), and (12) indicate that this occurs if $r_{dx} = 0$, $r_{dsx} = 0$, $U_{dw} = 0$, $U_d = V_d$.) The workers' excess optimism thus reduces the firm's safety precautions because workers are not willing to "buy" safety. If m is set at the optimum level, worker underestimation of risk leads to suboptimal safety.

An increase in the mandatory insurance benefit may decrease worker precautions. Mandatory insurance raises the amount of coverage, but leaves the insured worker in a position in which he perceives that the value of insurance protection is lower than the cost. The reduction in his perceived wealth will raise his precautions if the cost of precautions (in terms of utility) increases with income. On the other hand, the lower level of uninsured loss reduces the rewards for taking precautions. In the absence of income effects and without a change in firm precautions, an expansion of fair insurance coverage will generally reduce worker precautions, assuming that the firm cannot monitor precautions.[5]

Oi (p. 78) and Diamond (p. 81) are not necessarily correct in suggesting that workmen's compensation offers a remedy for suboptimal safety caused by excess worker optimism. An increase in m may or may not increase the firm's precautions. The ambiguity occurs because of two opposing influences on the firm's safety decision. An increase in m raises the firm's return to investment in safety by raising the cost of an accident, but it also increases the worker's desire to substitute wages for safety. The latter effect has not been considered in previous studies. It is possible that an increase in m to a level that would provide optimal insurance would reduce the level of safety as a result of the combined effect of moral hazard and the substitution of wages for safety.

2. *Worker Precautions Fixed*

In order to focus on the substitution of wages for safety following the introduction of mandatory insurance, it is assumed that workers' precautions are fixed. ($\partial d/\partial m = \partial d/\partial s = \partial d/\partial w = 0$). It is shown in Appendix B that an increase in the required level of benefits will *reduce* safety if $(1-r)/(1-a) \leqslant a/r(r_s/a_s)$. Since a/r will be nearly equal to one for small accident probabilities, this sufficient condition is likely to hold if the level of risk is underestimated more than r_s. The reduction in safety occurs because workers are less willing to sacrifice wages for safety when the uncompensated loss $(U-V)$ falls, and because workers will accept lower safety in order to partially offset the added insurance premium. The only opposing influence is the increased marginal cost of accidents for the employer because of the higher benefits.

Even without moral hazard, safety could fall as a result of an increase in mandatory insurance. This possibility has not been appreciated by other researchers. In fact, Diamond's proof (p. 81) that safety rises when benefits are increased contains an error.[6] The results here indicate that the exact nature of the underestimation of risk must be known before one can predict the response to mandatory insurance. In particular it must be determined whether the level

[5] From equations (9) and (11),

$$\frac{\partial d}{\partial m} - \frac{(1-a)}{a}\frac{\partial d}{\partial w} = \left[(1-r)V_{dm} - r_d V_m - \left(\frac{1-a}{a}\right) r U_{dw} - \left(\frac{1-a}{a}\right) r_d U_w\right] + D < 0$$

if $U_{dw} = V_{dm} = 0$.

[6] If $V_m > U_w$ (rather than $V_m < U_w$) his equation (25) has an indeterminant sign.

of risk is underestimated more than the marginal effect of the firm's precautions.

The crucial role of worker perceptions is illustrated by considering the situation in which d is fixed and the worker does not alter his estimate of risk in response to the firm's safety precautions ($r_s=0$). In this situation there will be no compensating wage differential. The firm will invest in safety until $C'=a_s(m+C)/a$, without regard to an effect on the wage rate. In this case, an increase in m will raise s and lower the wage rate. In the more general case in which workers respond to changes in safety ($r_s\neq 0$), workers may prefer to lower s following an increase in m.

It appears that a reduction in safety following the introduction of workmen's compensation is a distinct possibility. In fact James Chelius (1973, p. 63) finds statistically significant evidence that more generous workmen's compensation levels are associated with higher injury rates. This reduction could result from worker substitution of wages for safety or from moral hazard. Expected utility, evaluated using the actual probability, could fall if the reduction in safety is sufficiently large that it overwhelms the positive effect of insurance on utility. The change in utility evaluated at the actual risk level equals:

$$(19)\quad aU_w\left[\left(-\frac{C'}{a}+\frac{a_s(m+C)}{a^2}+a_sZ'\right)\times\left(1-\frac{r}{a}\left(\frac{a_s}{r_s}\right)\right)ds-\left(\frac{1-a}{a}\right)\left(1-\frac{V_m}{U_w}\right)dm\right]$$

If the firm is insured by the Workmen's Compensation Board or an insurance carrier, lack of perfect information on the firm's precautions[7] reduces safety and makes the coefficient of ds larger than shown in equation (19). This additional moral hazard increases the likelihood that expected utility will fall as a result of the mandatory insurance.

C. *Occupational Safety Regulations*

Governments have imposed many safety regulations (under OSHA, for instance) that would be undesirable in an economy with perfect information.[8] Worker underestimation of risk is often cited as an economic justification for the regulations, but there are at least three reasons why the regulations may not produce the desired results. First, it may be too expensive to regulate and monitor worker precautions. If d responds to economic incentives, safety could increase or decrease as a result of the regulation of s, depending on the parameters of $r(d,s)$ and $a(d,s)$. Second, if insurance benefits are at the market level (workmen's compensation does not constrain behavior), underestimation of risk could produce too much safety. In this case, legislated increases in safety precautions will clearly make workers worse off. Third, workers who underestimate risk will attempt to adjust their insurance coverage (if it is market determined) in response to the mandatory firm precautions. This substitution may reduce insurance and it may make the worker worse off.

With fixed worker precautions, it is shown in Appendix C that increases in mandatory firm precautions will reduce insurance coverage if $(1-r)/(1-a)\leqslant(a/r)(r_s/a_s)$.[9] This condition is identical to the condition that determined if mandatory insurance reduced safety. Mandatory safety will lower insurance coverage if workers underestimate the level of risk more than the marginal effect of firm safety. Since insurance might move further from the optimal level and safety levels may already be too high, the worker's expected utility could fall as a result of the regulation.

III. Conclusions

The model presented above indicates that information plays a crucial role in the insurance market and the market for safety. The efficiency of the market outcome de-

[7] Louise Russell examines the relationship between accident cost and firm premiums.

[8] See Smith.

[9] This is a sufficient condition.

pends on the exact nature of the imperfect information. For instance, if employee precautions are fixed, a decrease in the perceived level of risk unambiguously reduces market insurance below the optimal level only if the perceived marginal effect of employer precautions does not change. Under some assumptions, misperception leads to suboptimal insurance but excessive safety precautions.

Government policy may not be able to produce both optimal safety and optimal insurance when there is imperfect information. Moral hazard resulting from the difficulty of monitoring firm and employee precautions is present regardless of whether insurance is provided in the market or by a government insurance agency. The existence of worker underestimation of risk could justify mandatory insurance, but the level of safety may fall as a result of such a policy because workers will attempt to substitute wages for safer jobs. Moral hazard will also reduce the level of safety. Although misperception may have an adverse effect on the market level of insurance and safety, mandatory insurance or mandatory safety precautions may lower the utility of workers.

APPENDIX

A: MARKET DETERMINATION OF *m* AND *s*

1. *The General Model*

If worker perceptions are incorrect, competition for workers maximizes

$$U^*=r(d,s)U(w,d)+(1-r(d,s))V(m,d)$$
$$+\lambda\left[Z(a(d,s))-w-\left(\frac{(1-a(d,s))m+C(s)}{a(d,s)}\right)\right]$$

where $d=d(w,m,s)$ and $r_d(U-V)=-[rU_d+(1-r)V_d]$ from equation (8). The first-order conditions are

$$\frac{\partial U^*}{\partial w}=rU_w-\lambda+\lambda\left[\frac{\partial d}{\partial w}a_d\left(\frac{m+C}{a^2}+Z'\right)\right]=0$$

$$\frac{\partial U^*}{\partial m}=(1-r)V_m-\lambda\left(\frac{1-a}{a}\right)$$
$$+\lambda\left[\left(\frac{\partial d}{\partial m}\right)a_d\left(\frac{m+C}{a^2}+Z'\right)\right]=0$$

$$\frac{\partial U^*}{\partial s}=r_s(U-V)+\lambda\left[\frac{-C'}{a}\right.$$
$$+\left(\frac{\partial d}{\partial s}\right)a_d\left(\frac{(m+C)}{a^2}+Z'\right)$$
$$\left.+a_s\left(\frac{(m+C)}{a^2}+Z'\right)\right]=0$$

$$\frac{\partial U^*}{\partial \lambda}=Z-w-\frac{(1-a)m}{a}-\frac{C}{a}=0$$

Equations (15) and (16) are derived by substituting for λ.

2. *d Predetermined*

If *d* is no longer variable, the first-order conditions are

$$rU_w-\lambda=0$$

$$(1-r)V_m-\left(\frac{1-a}{a}\right)\lambda=0$$

$$r_s(U-V)+\lambda\left[\frac{-C'}{a}+\frac{a_s(m+C)}{a^2}+Z'a_s\right]=0$$

$$Z-w-\frac{(1-a)m}{a}-\frac{C(s)}{a}=0$$

Let

$$\frac{-C'}{a}+\frac{a_s(m+C)}{a^2}+Z'a_s=N$$

N must be negative if $\lambda>0$ and $(U-V)>0$.

Total differentiation of the first-order conditions gives

$$\begin{bmatrix} rU_{ww} & 0 & r_sU_w & -1 \\ 0 & (1-r)V_{mm} & -r_sV_m+\dfrac{\lambda a_s}{(a)^2} & \dfrac{-(1-a)}{a} \\ r_sU_w & -r_sV_m+\dfrac{\lambda a_s}{(a)^2} & B & N \\ -1 & \dfrac{-(1-a)}{a} & N & 0 \end{bmatrix} \begin{bmatrix} dw \\ dm \\ ds \\ d\lambda \end{bmatrix} = \begin{bmatrix} -U_w dx \\ +V_m dx \\ -r_{sx}(U-V)dx \\ 0 \end{bmatrix}$$

where B is assumed to be negative. The determinant $D<0$ if the second-order conditions hold and $D_{11}>0$ $D_{22}>0$ $D_{33}>0$.

If r is also a function of x, $r=r(d,s,x)$, $r_x=1$, and $r_{sx}=0$,

$$\frac{\partial w}{\partial x}=-U_w\frac{D_{11}}{D}+V_m\frac{D_{21}}{D}$$

$$\frac{\partial m}{\partial x}=-U_w\frac{D_{12}}{D}+V_m\frac{D_{22}}{D}$$

$$\frac{\partial s}{\partial x}=-U_w\frac{D_{13}}{D}+V_m\frac{D_{23}}{D}$$

where D_{ij} is the cofactor of row i and column j. Substituting $\lambda=rU_w$ one finds

$$D_{12}=\left(r_sU_w\left(\frac{1-a}{a}\right)-r_sV_m+\frac{rU_wa_s}{a^2}\right)N + \frac{(1-a)}{a}B$$

By substituting the values of U_w and V_m from the first-order conditions, it can be shown that

$$D_{12}<0 \text{ if } \left(\frac{1-r}{1-a}\right)\frac{a_s}{r_s}\geqslant\frac{a(2r-1)}{r}$$

since $N<0$, $r_{ss}<0$, and $B<0$ by assumption.

It can be shown that

$$\frac{\partial m}{\partial x}=-U_w\frac{D_{12}}{D}+V_m\frac{D_{22}}{D}<0$$

$$D_{13}=-(1-r)V_{mm}N-\frac{(1-a)}{a} \times\left[r_sU_w\left(\frac{-(1-a)}{a}\right)-r_sV_m+\frac{\lambda a_s}{(a)^2}\right]\gtrless 0$$

Substituting the values of U_w and V_m from the first-order conditions, it can be shown that

$$D_{13}<0 \text{ if } \left(\frac{1-r}{1-a}\right)\geqslant\left(\frac{r_s}{a_s}\right)\left(\frac{a}{r}\right)$$

$$D_{32}=rU_{ww}\left[\frac{-(1-a)}{a}\right]N-\left(\frac{1-a}{a}\right)r_sU_w -r_sV_m+\frac{\lambda a_s}{a^2}$$

$$D_{32}<0 \text{ if } \left(\frac{1-r}{1-a}\right)\leqslant\frac{r_s}{a_s}\frac{a}{r}$$

$$\frac{\partial s}{\partial x}=-U_w\frac{D_{13}}{D}+V_m\frac{D_{23}}{D}$$

The term $\partial s/\partial x$ will be positive if $(1-r)/(1-a)$ is sufficiently below $(r_s/a_s)(a/r)$ because $V_m(D_{23}/D)$ becomes more positive and $-U_w(D_{13}/D)$ becomes less negative (or positive) as $(1-r)/(1-a)$ falls below $(r_s/a_s)(a/r)$. Alternatively one can write out the expression as

$$\frac{\partial s}{\partial x}=\frac{1}{D}\left[U_w(1-r)V_{mm}-rV_mU_{ww}\left(\frac{1-a}{a}\right)\right]N +\left[V_m+\frac{(1-a)}{a}U_w\right] \times\left[-r_sU_w\frac{(1-a)}{a}-r_sV_m+\frac{\lambda a_s}{(a)^2}\right]\frac{1}{D}\gtrless 0$$

The first term is positive if $V_{mm}/U_{ww} < (V_m/U_w)^2$, a plausible relationship since $m \leqslant w$. The second term is positive if $(1-r)/(1-a) < (r_s/a_s)(a/r)$.

$$\frac{\partial w}{\partial x} = -U_w \frac{D_{11}}{D} + V_m \frac{D_{21}}{D}$$

$$\therefore \frac{\partial w}{\partial x} > 0 \text{ if } \left(\frac{1-r}{1-a}\right)\frac{a_s}{r_s} \geqslant \frac{a(2r-1)}{r}$$

If $r_x = 0$ and $r_{sx} < 0$,

$$\frac{\partial s}{\partial x} = -\frac{D_{33}}{D}(U-V)r_{sx} < 0$$

$$\frac{\partial m}{\partial x} = -\frac{D_{32}}{D}(U-V)r_{sx} \lesseqgtr 0$$

$$\frac{\partial m}{\partial x} > 0 \text{ if } \left(\frac{1-r}{1-a}\right) \leqslant \frac{r_s}{a_s}\frac{a}{r}$$

$$\frac{\partial w}{\partial x} = -\frac{D_{31}}{D}(U-V)r_{sx} \lesseqgtr 0$$

$$\frac{\partial w}{\partial x} > 0 \text{ if } \left(\frac{1-r}{1-a}\right) \geqslant \frac{r_s}{a_s}\frac{a}{r}$$

B: *M* Determined by Government

1. *d Endogenous*

If m is exogenous, the first-order conditions in the general model reduce to

$$rU_w + \lambda\left[-1 + \left(\frac{\partial d}{\partial w}\right)a_d\left(\frac{m+C}{a^2} + Z'\right)\right] = 0$$

$$r_s(U-V) + \lambda\left[\frac{-C'}{a} + \left(\frac{\partial d}{\partial s}\right)a_d\left(\frac{(m+C)}{a^2} + Z'\right)\right.$$
$$\left. + a_s\left(\frac{(m+C)}{a^2} + Z'\right)\right] = 0$$

$$Z - w - \frac{(1-a)m}{a} - \frac{C(s)}{a} = 0$$

It can easily be shown that when $r_{dx} = r_{sx} = r_{ds} = U_{dw} = 0$, and $U_d = V_d$, $\partial w/\partial x = -U_w \times (D_{11}/D)r_x > 0$, where D_{11} is a cofactor of the matrix that results from total differentiation of the first-order conditions. From the budget constraint $\partial s/\partial x < 0$, since $(\partial d/\partial w)a_d(m + C/a^2) + Z') < 1$.

Consider a change in x such that $r_x = 0$ and $r_{sx} < 0$:

$$\frac{\partial s}{\partial x}, (r_x = 0, r_{sx} < 0) = -(U-V)r_{sx}\frac{D_{22}}{D} < 0$$

From the budget constraint it follows that

$$\frac{\partial w}{\partial x}, (r_x = 0, r_{sx} < 0) > 0$$

2. *d Predetermined*

If d is fixed, $\partial d/\partial w = \partial d/\partial s = 0$ and the first-order conditions are

$$rU_w - \lambda = 0$$

$$r_s(U-V) + \lambda N = 0$$

$$Z - w - \frac{(1-a)m}{a} - \frac{C(s)}{a} = 0$$

Total differentiation gives

$$\begin{bmatrix} rU_{ww} & r_sU_w & -1 \\ r_sU_w & B & N \\ -1 & N & 0 \end{bmatrix} \begin{bmatrix} dw \\ ds \\ d\lambda \end{bmatrix}$$

$$= \begin{bmatrix} -U_w dx \\ -(U-V)r_{sx}dx + r_sV_m dm - \lambda\frac{a_s}{(a)^2}dm \\ \frac{(1-a)}{a}dm \end{bmatrix}$$

As shown above $\partial w/\partial x > 0$ and $\partial s/\partial x < 0$ if $r_{sx} \leqslant 0$

$$\frac{\partial s}{\partial m} = \left(r_sV_m - \frac{\lambda a_s}{a^2}\right)\frac{D_{22}}{D} + \left(\frac{1-a}{a}\right)\frac{D_{23}}{D}$$

$$= \left[r_sV_m - \frac{\lambda a_s}{(a)^2}\right]\left[\frac{-1}{D}\right]$$

$$-\frac{(1-a)}{a}\left[rU_{ww}N+r_sU_w\right]\left[\frac{1}{D}\right]$$

$$\frac{\partial s}{\partial m}=\left[-r_sV_m+\frac{rU_w a_s}{(a)^2}-\frac{(1-a)}{a}r_sU_w\right]\left(\frac{1}{D}\right)$$

$$-\left(\frac{1-a}{a}\right)(rU_{ww}N)\left(\frac{1}{D}\right)$$

At the market levels of m and s, $\partial s/\partial m<0$ if $(1-r/1-a)\leqslant(r_s/a_s)a/r$ since $D>0$, $U_{ww}<0$, and $N<0$.

$$\frac{\partial w}{\partial m}=\left[r_sV_m-\frac{\lambda a_s}{(a)^2}\right]\left(\frac{D_{21}}{D}\right)+\left(\frac{1-a}{a}\right)\left(\frac{D_{31}}{D}\right)$$

$$\frac{\partial w}{\partial m}=N(r_sV_m)\left[-1+\frac{1-r}{r}+\left(\frac{1-r}{1-a}\right)\frac{a_s}{r_s}\frac{1}{a}\right]$$

$$\times\frac{1}{D}+\left(\frac{1-a}{a}\right)[B]\frac{1}{D}$$

$$\frac{\partial w}{\partial m}<0 \text{ if } \frac{(1-r)}{(1-a)}\frac{a_s}{r_s}\geqslant\left(\frac{2r-1}{r}\right)a$$

3. *Change in Utility following Mandatory Insurance*

Evaluate the expected utility with the actual probabilities

$$U^*=aU(w)+(1-a)V(m)$$

$$dU^*=(U-V)a_s\,ds+aU_w\,dw+(1-a)V_m\,dm$$

If we evaluate dU^* at the market levels

$$r_s(U-V)=-\lambda\left[\frac{-C'}{a}+\frac{a_s(m+C)}{a^2}+a_sZ'\right]$$

and $\lambda=aV_m(1-r)/(1-a)=rU_w$ from the first-order conditions. From the budget constraint

$$dw=\left[\frac{-C'}{a}+\frac{a_s(m+C)}{a^2}+a_sZ'\right]ds$$

$$-\left(\frac{1-a}{a}\right)dm$$

substituting dw, $U-V$, and λ gives equation 20.

C: *s* Regulated, *m* Unregulated, *d* Predetermined

If s is exogenous and d is fixed, the first-order conditions are

$$rU_w-\lambda=0$$

$$(1-r)V_m-\lambda\left(\frac{1-a}{a}\right)=0$$

$$Z-w-\frac{(1-a)m}{a}-\frac{C(s)}{a}=0$$

Total differentiation, with $r_x=1$, gives

$$\begin{bmatrix} rU_{ww} & 0 & -1 \\ 0 & (1-r)V_{mm} & \dfrac{-(1-a)}{a} \\ -1 & \dfrac{-(1-a)}{a} & 0 \end{bmatrix}\begin{bmatrix} dw \\ dm \\ d\lambda \end{bmatrix}$$

$$=\begin{bmatrix} -U_w\,dx-r_sU_w\,ds \\ +V_m\,dx+r_sV_m\,ds-\dfrac{\lambda a_s}{(a)^2}ds \\ -N\,ds \end{bmatrix}$$

where $N<0$ (see above).

$$\frac{\partial m}{\partial x}=-U_w\frac{D_{12}}{D}+V_m\frac{D_{22}}{D}$$

$$\frac{\partial w}{\partial x}=-U_w\frac{D_{11}}{D}+V_m\frac{D_{21}}{D}$$

$$D_{12}=\frac{1-a}{a}>0,\ D_{22}<0,\ D_{11}<0,\ D>0$$

$$\frac{\partial m}{\partial x}<0, \text{ and } \frac{\partial w}{\partial x}>0$$

$$\frac{\partial m}{\partial s}=-r_sU_w\frac{D_{12}}{D}+\left(r_sV_m-\frac{\lambda a_s}{(a)^2}\right)\left(\frac{D_{22}}{D}\right)$$

$$-N\left(\frac{D_{32}}{D}\right)$$

$$D_{32} = rU_{ww}\frac{(1-a)}{a} < 0 \quad D_{11} = -\left(\frac{1-a}{a}\right)^2 < 0$$

$$D_{31} = (1-r)V_{mm} < 0 \quad D_{22} = -1 < 0$$

$$D_{12} = \frac{1-a}{a} > 0$$

$$\frac{\partial m}{\partial s} = \left[-r_s U_w\left(\frac{1-a}{a}\right) - \left(r_s V_m - \frac{\lambda a_s}{(a)^2}\right) - NrU_{ww}\left(\frac{1-a}{a}\right)\right]\frac{1}{D} \gtrless 0$$

If s is increased above the market level,

$$\frac{\partial m}{\partial s} < 0 \text{ if} \left(\frac{1-r}{1-a}\right) \leqslant \frac{a}{r}\left(\frac{r_s}{a_s}\right)$$

$$\frac{\partial w}{\partial s} = -r_s U_w \frac{D_{11}}{D} + \left(r_s V_m - \frac{\lambda a_s}{(a)^2}\right)\frac{D_{21}}{D} - N\frac{D_{31}}{D}$$

$$\frac{\partial w}{\partial s} = \left[r_s U_w\left(\frac{1-a}{a}\right)^2 + \left(r_s V_m - \frac{\lambda a_s}{(a)^2}\right)\left(\frac{1-a}{a}\right) - N(1-r)V_{mm}\right]\frac{1}{D}$$

$$\frac{\partial w}{\partial s} > 0 \text{ if} \left(\frac{1-r}{1-a}\right) \geqslant \frac{a}{r}\left(\frac{r_s}{a_s}\right)$$

REFERENCES

J. R. Chelius, "An Empirical Analysis of Safety Regulations," in *Supplemental Studies*, Vol. III, U.S. National Commission on State Workmen's Compensation Laws, Washington 1973, 53–66.

______, *Workplace Safety and Health: The Role of Workers' Compensation*, Washington 1977.

P. Diamond, "Insurance Theoretic Aspects of Worker's Compensation," in Alan S. Blinder and Philip Friedman, eds., *Natural Resources, Uncertainty, and General Equilibrium Systems*, New York 1977.

I. Ehrlich and G. S. Becker, "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*, 623–48.

D. Epple and A. Raviv, "Product Safety: Liability Rules, Market Structure, and Imperfect Information," *Amer. Econ. Rev.*, Mar. 1978, *68*, 80–95.

P. Gregory and M. Gisser, "Theoretical Aspects of Workmen's Compensation," in *Supplemental Studies*, Vol. I, U.S. National Commission on State Workmen's Compensation Laws, Washington 1973, 108–28.

Howard Kunreuther et al., *Disaster Insurance Protection: Public Policy Lessons*, New York 1978.

A. L. Nichols and R. Zeckhauser, "Government Comes to the Workplace: An Assessment of OSHA," *Publ. Interest*, Fall 1977, *49*, 39–69.

W. Oi, "Workmen's Compensation and Industrial Safety," in *Supplemental Studies*, Vol. I, U.S. National Commission on State Workmen's Compensation Laws, Washington 1973, 42–107.

M. V. Pauly, "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection," *Quart. J. Econ.*, Feb. 1974, 88, 44–62.

S. Peltzman, "The Effects of Automobile Safety Regulation," *J. Polit. Econ.*, Aug. 1975, *83*, 677–725.

M. Rothschild and J. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quart. J. Econ.*, Nov. 1976, *90*, 629–50.

L. B. Russell, "Safety Incentives in Workmen's Compensation Insurance," *J. Human Resources*, Summer 1974, *9*, 361–75.

S. Shavell, "On Moral Hazard and Insurance," *J. Quart. Econ.*, Nov. 1979, *93*, 541–62.

Adam Smith, *The Wealth of Nations*, New York 1937.

Robert S. Smith, *The Occupational Safety and Health Act*, Washington 1976.

M. Spence and R. Zeckhauser, "Insurance, Information, and Individual Action," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 380–87.

M. Spence, "Consumer Misperceptions, Product Failure and Producer Liability," *Rev. Econ. Stud.*, Oct. 1977, *44*, 561–72.

R. Thaler and S. Rosen, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor E. Terlecky, ed., *Household Production and Consumption*, New York 1975, 265–301.

A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science*, Sept. 27, 1974, *185*, 1124–31.

______, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, Mar. 1979, *47*, 263–91.

W. K. Viscusi, "Wealth Effects and Earnings Premiums for Job Hazards," *Rev. Econ. Statist.*, Aug. 1978, *60*, 408–16.

O. E. Williamson et al., "Externalities, Insurance and Disability Analysis," *Economica*, Aug. 1967, *24*, 235–53.

C. Wilson, "A Model of Insurance Markets with Incomplete Information," *J. Econ. Theory*, Oct. 1977, *16*, 167–207.

Black-White Human Capital Differences: Impact on Agricultural Productivity in the U.S. South

By WALLACE E. HUFFMAN*

In a dynamic environment, schooling of farmers and agricultural extension have the potential for enhancing the efficiency of agricultural production. In the U.S. South, a relatively large number of blacks have operated farms since emancipation. In the segregated school systems, these black farmers obtained lower quality and fewer years of schooling than white farmers. The public sector extension service had the potential for mitigating the effects of lower quality black education on farm production efficiency. In eleven of the sixteen southern states, however, the Extension Service was completely segregated, and the services provided to black farmers were fewer and seem to have been of lower quality than those provided white farmers.

The objective of this study is to present econometric estimates of productivity differences on black and white operator farms in the U.S. South. The results from fitting a production function to county data for 1964 show that the quantity and quality of farmers' education and of extension are the primary sources of differential productivity on black and white farms. The lower productivity of black farms is undoubtedly one of the factors contributing to the exodus of black farmers from southern agriculture at double the rate of white farmers during the 1950's and 1960's when agricultural technology was changing rapidly.

The paper is organized as follows. Section I discusses sources of managerial skill differences. A model for investigating productivity differences on black and white operator farms is presented in Section II. In Section III, the empirical measures of the variables and the estimate of the production function are presented and discussed. The last section contains the implications and conclusions.

I. The Sources of Managerial Skill Differences of Black and White Farmers in the U.S. South

In a technically and economically dynamic environment, schooling of farmers and agricultural information have the potential for enhancing the efficiency of agricultural production (see Finis Welch, 1970, and my 1977 paper). Many adjustments in farming are required when new and potentially better opportunities become available. These opportunities may arise because of changes in market conditions caused by shifts in demand for farm output, by unexpected changes in environmental variables affecting production, and by the development of new technology that changes the potential nature of supply. Farmers differ in their ability to respond to these changes, and if managerial skill differs by race of farmer, it may be an important source of comparative advantage of one group over another.

A. *Training*

In this study, training that may enhance the managerial ability of farmers, and hence be a source of differential ability between races, is the quantity and quality of schooling and past farming experience. In 1964, the only year that data are available, the *Census of Agriculture, 1964* shows large differences in the years of schooling completed by black and white farm operators in the U.S. South. Nonwhite farm operators had completed only 5 years of schooling, but

*Associate professor of economics, Iowa State University. T. W. Schultz, Joe Reid, Jr., Richard Perrin, J. Peter Mattila, George Borts, and an anonymous reviewer provided helpful suggestions on earlier drafts. Journal paper no. J-9776 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 2078.

white operators had completed 9.5 years. The farmers of 1964 were born largely between 1900 and 1944, and attended formal schooling between 1906 and 1958.

Although the exact role of schooling quality in later managerial performance is unknown, the quality of schooling of black farmers seems to have been inferior to the schooling quality of southern white farmers. Differences exist at both the preschool and formal schooling level. Black children of this era were generally handicapped relative to white children in the U.S. South before entering school because of the generally low levels of completed schooling and of literacy of their parents. Because of slavery (before 1865) and the legislated discrimination against the schooling of blacks in the South, they got started slowly relative to whites in obtaining schooling.[1] Blacks born in the early 1900's were only one or two generations away from slavery, and only modest progress had been made in financing schooling for blacks in the South between 1865 and 1900 (see Welch, 1973b). The lower schooling levels of black parents reduced the potential for teaching their children basic skills and discipline before entering school, and also for assisting their children with homework after entering school. One effect of this differing family background is that black children were less well prepared for formal schooling than were white children. One piece of evidence is the relatively high retention rates in first grade of students in black schools. Between 1910 and 1940, the ratio of enrollment in first to enrollment in second grade was about 2, suggesting that each child spent twice as long in first grade as in second grade. For all *U.S.* schools, it was about 1.5 (Welch 1973a,b).

Between 1900 and 1940, the differences in characteristics of black and white schools were such that they suggest large quality differences in the South. In 1896, the U.S. Supreme Court sanctioned the "separate-but-equal" schools for whites and for blacks. This removed the legal and social pressure that previously existed for equality. Although there was a persistent upward trend in average days of school attended by students in black schools after 1900, relative differences in daily attendance between black schools and southern white schools widened and then moved toward equality. Between 1900 and 1940, teachers' salaries in black schools were approximately one-half as large, and annual per pupil expenditures were approximately one-third as large for black schools as for southern white schools (see Welch, 1973b). Number of pupils enrolled per classroom teacher was about 1.5 times larger in black schools than in all *U.S.* schools. After about 1940, but before the 1954 Supreme Court decision against separate-but-equal schools, Welch (1973a) concludes that relative differences between black and white schools were steadily decreasing.

In farming, intergeneration transfer of information may be important, especially during periods when the environment is technically static and "rule of thumb" decision making performs well. Fathers may pass on useful information about planning, managing, and financing a farm business to their sons (and daughters) as they work and learn on their father's farm. Even for this training, black farmers were disadvantaged relative to most white farmers. The reason is that only a few generations of black farmers had the opportunity of independent farming experience where they made and bore the financial consequences of farm management and marketing decisions, and of credit and long-term debt decisions.[2] Thus, the lower quality training of black farmers could be expected to affect their ability to compete with white farmers in producing agricultural output. Their disadvantage might be mitigated, however, if they had access to superior information and agricultural technology specifically designed for their type and size of farms and their decision-making skill level.

[1] Robert Fogel and Stanley Engerman (pp. 39–40) indicated that in 1850, 73 percent of male slaves were unskilled farm fieldhands, and education for them was considered unnecessary. Only 7 percent of slaves held managerial positions.

[2] Black sharecroppers did have the opportunity to learn from white landowners (see Joseph Reid, 1977).

B. *The Organization of Public Agricultural Research and Extension*

Land grant colleges and the Extension Service are the major public sector sources of agricultural research, and of practical and timely information for farmers. Most of the research is conducted by the agricultural experiment stations. Although a few states established stations on their own during the 1870's and 1880's, the Hatch Act (1887) authorized federal support to each state that would establish an agricultural experiment station in connection with its land-grant college. This Act established agricultural experiment stations in each of the states. In the early years, all of the funding for the stations was federal, but over time state matching of funds was required, and now federal support of agricultural experiment station research is only 30 percent of the total budget.

Although much of the work in the agricultural experiment stations in early years served to facilitate the transfer and adoption of techniques developed by farmers and farm machinery manufacturers, agricultural research in later years has produced increments to basic knowledge and applied research. Some of the applied research attempts to increase agricultural output (for example, new or improved crop varieties, decision-making aids and schemes, and final agricultural products) while others attempt to maintain previous technological gains. The performance of much of agricultural technology is sensitive to local environmental factors and resource endowments, including size of farm and managerial skill of farmers. Thus, widespread direct interstate borrowing of applied research products is generally limited, and intrastate research must be targeted to the needs of different locations and types of farms.

Studies by Zvi Griliches (1964) and Robert Evenson (1971, 1980) have shown that public sector investments in agricultural research (and extension) have increased the productivity of *U.S.* agriculture. All farmers, however, inherently do not have equal access to new technology. Operators of large farms have a greater incentive to search and experiment than do operators of small farms (see my 1977 paper). Farmers in different geoclimatic regions may have differential access because of technological-environmental interactions (see Griliches, 1957; Evenson, 1980). Some technology may be profitable only when applied on a large scale. Finally, some operators may have more skill for acquiring and interpreting information, and are thereby better able to experiment, sort out relevant facts, and make modifications for their farming situation.

The Extension Service has the potential to be a substitute for high managerial skill of farmers. The Extension Service, established with federal-state coordination in 1914, is the most important public sector source of information to farmers, but it is only one of many private and public institutions providing information to them. Agricultural extension personnel assemble, organize, and interpret market information, simplify technical information, and develop resource management schemes for disseminating to farmers. They also demonstrate new farming techniques and consult directly with farmers on specific production and management problems. Tough problems are to be referred to state extension specialists or to experiment station researchers. Thus, the Extension Service has attempted to develop an information system that enhances information transfer and adoption of new technology by linking farmers to the expertise of state extension specialists and to researchers at experiment stations.

In the U.S. South, the organization of agricultural research and extension seems to have contributed to unequal access to new technology by black and white farmers. The original land-grant colleges, established by the Morrill Act of 1862, developed as segregated institutions for whites, and the Land-Grant Act of 1890 authorized the establishment of separate-but-equal land-grant colleges for blacks. All sixteen southern states and Missouri established "Colleges of 1890" under this Act.[3] State and federal financial support, especially for

[3] Edward Eddy (p. 291) presents a list of the seventeen land-grant colleges of 1890. During the early years of these colleges, most of their students were enrolled in courses at the elementary and high school level because few blacks had completed high school.

TABLE 1—EXPENDITURES AND STATE AND FEDERAL APPROPRIATIONS FOR THE SEVENTEEN WHITE (1862) AND BLACK (1890) LAND-GRANT COLLEGES IN THE U.S. SOUTH, 1945–60
(Thousands of Current Dollars)

	1945		1950		1955		1960[a]	
Item	White	Black	White	Black	White	Black	White	Black
Total Expenses for Educational and General Purposes	55,942	4,302	128,858	13,072	188,828	20,191	306,664	27,102
a) State Government Appropriations	24,920	3,088	67,614	9,994	107,108	15,852	169,137	22,144
b) Funds of Federal Origin	24,287	519	43,220	2,344	36,477	627	67,895	854
Total Expenditures on Organized Research	9,872	1	19,367	16	37,259	31	70,460	115
a) Regular Federal Land-Grant Appropriations for Research (Experiment Station)	1,571	0	4,398	0	7,604	0	11,920	0
Total Expenditures on Extension and Public Information	20,107	88	31,473	282	44,447	214	66,443	488
a) Regular Federal Land Grant Appropriations for Cooperative Extension	10,473	0	15,741	0	19,861	0	26,026	47

Sources: U.S. Office of Education, 1947, 1951, 1956, 1961.
[a]For sixteen states—West Virginia did not have a black land-grant college in 1960.

agricultural research and extension, has been extremely unequal for the 1862 and 1890 colleges (see Table 1). For the period 1945–60, the 1862 land-grant colleges made more than 99 percent of the total expenditures on organized research by southern white and black land-grant colleges, and they received all the regular federal appropriations for agricultural experiment station research. The decision on allocating federal experiment station funds between land-grant colleges was made by each state's legislature or by its governor, but in every southern state, all of the federal funds for agricultural experiment stations were allocated to the 1862 (white) land-grant institutions. Furthermore, none of the black land-grant college researchers had direct access to an on-campus experiment station, except in Texas where a branch station was located at Prairie View A and M. Although black farmers have had about 5 percent (1945–60) of agricultural sales in the South, the black land-grant colleges have had few research resources for developing new agricultural technology specifically designed for small, low-skill, limited-resource black farmers. Given the political reality of obtaining state support for experiment station research in the South, the agricultural experiment stations of the 1862 colleges undoubtedly targeted applied research to the needs of white rather than to the needs of black farmers.[4]

The Extension Service had the potential to mitigate the effects of low skill levels of black farmers on differential access to new technology and on farm management, but this potential was not realized. The Morrill Act of 1890 required that state legislatures designate either an 1862 or an 1890 land-grant college to administer the extension of information to rural people. In the seventeen states with segregated land-grant col-

[4]A sizeable percentage of black farmers have been crop-share tenants (57 percent in 1940 and 40 percent in 1960). Thus, some of them may have benefited from applied research targeted to their white landlords. Others were undoubtedly made worse off by new technology that reduced the demand for farm labor and crop-share tenants (see Richard Day).

leges, the white land-grant college was chosen to administer the total extension program, but over time, a segregated structure developed. In eleven of the southern states where most of the rural blacks lived, the Extension Service was segregated from the state offices down to the local level and, the services provided to black farmers seem to have been inferior to those provided to white farmers.[5]

The offices of the state staff for the white Extension Service, located at the white land-grant colleges, were well staffed with generally well-trained specialists in a large number of subject areas (see U.S. Commission on Civil Rights (USCCR), 1965, p. 26). In 1960, the number of agricultural and home economics extension staff members operating at or from white land-grant colleges in the eleven southern states with completely segregated systems was 896 or 1 for each 1,186 white farmers (see Office of Education (USOE), 1961). These extension personnel also had direct access to the researchers of the state experiment stations.

In contrast, the offices of the state staff of the black Extension Service were located at the black state land-grant colleges, except in Mississippi where they were located in Jackson and not associated with a college, in Alabama where they were located at the private black Tuskegee Institute, and in Arkansas where they were located at the white land-grant college in Fayetteville. The black land-grant colleges had a small budget for extension and public information (see Table 1), few well-trained specialists, and a small extension staff (see USCCR, 1965, pp. 25–26). In 1960, eight states had offices of the state staff located at black land-grant colleges. For these eight states, the number of agricultural and home economics extension staff members operating at or from black land-grant colleges and universities was 74 or 1 for each 2,156 black farms, compared with one white state staff member for each 1,123 white farms in these states (see USOE, 1961). These black state extension staffs had little direct research support. Furthermore, as late as the early 1960's, regular contact between the state black and white extension personnel in the eleven southern states with segregated Extension Services did not exist, except in North Carolina, Mississippi and Texas (see USCCR, 1965).

In counties where both the black and white extension services had personnel, they had separate offices, supporting staffs, and equipment, and their personnel had different training. Agents were segregated for extension training, except in North Carolina, and the training of white agents was longer, more comprehensive, and more detailed than for black agents (see USCCR, 1965, pp. 30–36). In some counties that had a large number of black farm families, there was no black extension personnel. They were not assigned to counties where strong sentiment against such action existed.[6] Furthermore, the white county extension personnel seem to have provided only minimal assistance to black farmers (see USCCR, 1965, p. 40) in these counties.[7] Although good data do not exist, the size of black county extension staff was small relative to the size of its potential audience and compared to the potential audience of the white Extension Service.[8] Thus, it is clear that extension and schooling have been of lower quality and less available to black than to white farmers.[9] But we

[5] These states were Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, Virginia, and Arkansas. In 1960, these eleven states had 97 percent of the black farm operators in the sixteen southern states.

[6] The placement of black extension personnel in a county was determined by availability of federal, state, and county funds, the size of the black rural population, and the willingness of a county to have black extension workers.

[7] For meeting local needs, federal extension officials believe that local people must help prepare annual county extension plans. Blacks were involved in making plans only in counties where black extension personnel were present. No attempt was made in other counties to include blacks locally (see USCCR, 1965).

[8] One piece of evidence shows the order of magnitude of the inequality of black-white extension funding. For 1925-37, black farmers operated about 27 percent of all farms in the U.S. South, but expenditures for extension for blacks was roughly 6 percent of total (federal, state, and county) funds allocated for agricultural extension work (see Office of Education, 1949, p. 28).

[9] Civil Rights audits in 1969 of state Extension Services in the U.S. South show that pervasive discrimination in distribution of services to black and

would like to know how these differences translate into effects on agricultural productivity in the U.S. South.

II. A Model for Investigating Productivity Differences on Black and White Operator Farms

An aggregate production function provides the framework for quantifying the economic significance of differences in quantity and quality of black and white operator schooling and extension on agricultural productivity. The specification of the production function explicitly parameterizes potential productivity differences between races, and input quantities and estimates of the parameters are used to quantify these productivity differences.

The production function is

$$(1)\quad ln\, Y=\alpha_0+(\alpha_1+\beta_1\rho) ln\, MACH + (\alpha_2+\beta_2\rho) ln\, LIVST + (\alpha_3+\beta_3\rho) ln\, FS + (\alpha_4+\beta_4\rho) ln\, A + (\alpha_5+\beta_5\rho) ln\, L + \alpha_6 ln\, E + \alpha_7 ln\, X + \sum_k d_k R_k$$

where

ρ = share of livestock products in total farm output

$MACH$ = aggregate machinery input

$LIVST$ = aggregate livestock input

FS = aggregate fertilizer and seed input

A = aggregate composite farmland input:

A $=\Sigma_i \Sigma_j a_{ij} A_{ij}$, A_{ij} = aggregate acres, $i=1$ for white operators, 2 for black operators, $j=1$ for cropland, 2 for noncropland

FA $=\Sigma_i \Sigma_j A_{ij}$ = aggregate acres of farmland

L = aggregate composite operator and hired farm labor input:

L $=\Sigma_i l_i L_i + hH$, L = aggregate composite farm labor input

L_i = aggregate days operator labor; H = aggregate days of hired farm labor

LAB $=\Sigma_i L_i + H$ = aggregate days of operator and hired labor

E = aggregate composite education index of farm operators:

E $=\Sigma_i e_i E_i$, E_i = years of school completed

ED $=\Sigma_i E_i$ = aggregate years of schooling completed, all farm operators

X = aggregate composite agricultural extension index:

$X=\Sigma_i f_i X_i$, X_i = days of extension input, $i=1$ for white extension, 2 for black extension

EXT $=\Sigma_i X_i$ = aggregate days of agricultural extension input

R_k = regional dummy variables

The function is an extension of the Cobb-Douglas where input coefficients are a linear function of the mix of output ρ, measured as the livestock output share of total farm output. This functional specification permits the input-output relationship to vary by farm product mix and thereby to better fit observations differing widely in crop-livestock mix of output (see Griliches, 1963; my 1976 paper).[10]

Composite inputs, consisting of a component for white operator and for black operator farms, are hypothesized where productivity of inputs might be expected to differ by race, and where data are available on input usage by race of farm operator. The coefficients of the components of the composite inputs permit differential weighting due to productivity differences. The production function is, however, a non-linear function in the unknown parameters of the composite land, labor, education, and extension inputs. The method applied here to linearize these indexes is an approximation by

white farmers continues to be a problem (see USCCR, 1973).

[10] Discussions can be found elsewhere on potential problems with existence of aggregate production functions (Franklin Fisher; Robert Hall; John H. A. Green), on statistical identification of the production function (Griliches and Vidar Ringstad), on simultaneous equation bias (Irving Hoch; Arnold Zellner), and on the importance of land tenure arrangements (Stephen DeCanio; Reid, 1976, 1977).

Taylor-series expansion, ignoring second- and higher-order terms, about the equal productivity loci.[11]

Linearization of farmland, labor, education, and extension inputs creates a set of inputs that can be easily constructed from available data. Each of the linearized functions enters equation (1) as a function of a simple summation of the unadjusted components of the composite input (for example, for land, it is *ln* (FA)) and a ratio formed by dividing individual components by the simple aggregate input (for example, for land, the ratios are $(A_{11}+A_{21})/FA$, A_{21}/FA, and A_{22}/FA).[12] For education and extension the parameterization is $ln\, ED+[(e_2-e_1)/e_1^0]E_2/ED$ and $ln\, EXT+[(f_2-f_1)/f_1^0]X_2/EXT$, respectively, where the coefficient of E_2/ED is the relative difference in the productivity of a year of black operator schooling compared with a year of white operator schooling and the coefficient of X_2/EXT is the relative difference in the productivity of a unit of black extension compared with a unit of white extension. Thus, we expect the estimated coefficients of E_2/ED and X_2/EXT to be negative and significantly different from zero, if a unit of black operators' schooling (black extension) is less productive than a unit of white operators' schooling (white extension).

III. The Empirical Analysis

This section discusses the data set, the measurement of the variables, and the estimated aggregate production function. Investigating productivity differences of black and white operator farms would be facilitated if separate data on inputs and outputs by race of operator were available. Although the *Census of Agriculture, 1964* provides state level data that can be used to derive separate inputs and outputs for white and for nonwhite operator farms in fifteen southern states, these data do not provide enough observations for fitting production functions. Thus, these state level data are useful primarily as descriptive information. At the county level, the *Census of Agriculture, 1964* provides only partial information on the separate characteristics of white and nonwhite operator farms in the U.S. South. The data base is obtained by combining these county data with unpublished U.S. Department of Agriculture (USDA) data on the white and black Extension Services and information from USDA publications of the same period.[13]

The observations are county aggregates for the 295 counties of North Carolina, South Carolina, Mississippi, and Alabama. These four states in the U.S. South were chosen because they had the largest number of black farm operators in 1964; they had 58 percent of all black farm operators in the sixteen southern states (and 97.5 percent of the nonwhite farmers were black). These states also represent different parts of the South, the Mid South, and the Deep South.

A. *Empirical Measures of the Variables*

The derivation of key variables is presented to aid in assessing the empirical results. Farm output is measured as the value of all farm products sold, crops plus live-

[11] DeCanio has used a similar model for investigating productivity differences of black and white operator farms in the postbellum South.

[12] For land, the approximation is $ln\, A \sim c_0 + ln\, FA + c_1(A_{11}+A_{21})/FA + c_2(A_{21}/FA)+c_3(A_{22}/FA)$, where the unknown parameters $c_1=(a_{11}-a_{12})/a_{11}^0$, $c_2=(a_{21}-a_{11})/a_{11}^0$, and $c_3=(a_{22}-a_{12})/a_{11}^0$ show the relative difference in productivity of an acre of (a) cropland compared with an acre of other farmland, (b) cropland on black operator farms compared with cropland on white operator farms, and (c) other farmland on black operator farms compared with other farmland on white operator farms, respectively. For labor, the approximation is

$$ln\, L \sim \gamma_0 + ln\, LAB + \frac{(l_1-h)}{l_1^0}\frac{L_1+L_2}{LAB} + \frac{(l_2-l_1)}{l_1^0}\frac{L_2}{LAB}$$

where the estimate of $(l_2-l_1)/l_1^0$ shows the relative difference in productivity between black and white operator labor.

[13] One might ask what is unique about 1964? It is the year in which the Civil Rights Act was passed. One effect of this Act was to make illegal a separate black and white Extension Service. Also, the *Census of Agriculture, 1964* is the only one to present data on years of schooling completed by farm operators.

stock, and livestock products.[14] The share of livestock products in total farm output is measured as the sales of livestock and livestock products divided by total farm output. Farmland is defined in this study as cropland harvested and nonwoodland pasture land. This definition excludes land in farms that are relatively unproductive, for example, idle cropland, woodland, and wasteland.[15] Total acres of farmland and acres of cropland (harvested) on white and black operator farms are reported in U.S. Bureau of the Census (1967).

The input of farmlabor services is derived from data on hours worked and expenditure data, and it is measured as annual man-days of farm work. Average annual days of farm work per farm operator in a county are estimated as the state average days of farm work by all farmers (see U.S. Bureau of the Census, 1968), less the net difference between the state and the county average days of off-farm work per farm operator (see U.S. Bureau of the Census, 1967). Separate county data on days of farm work by race of operator are not available in the *Census*. A measure of total days of farm work by black operators in a county was obtained by multiplying the above average days of farm work by all operators by the number of black operators.[16] Days of hired labor are derived as annual expenditure on hired labor divided by the state average daily wage rate in 1964 for hired farmlabor (see USDA, 1965). To obtain the total days of operator and hired farmlabor, the days of hired farmlabor were multiplied by 0.872 in Alabama, North Carolina, and South Carolina, and by 0.923 in Mississippi to adjust for differences in average length of work day (see Walter Sellers) and added to days of operator labor.

The aggregate education level of all farm operators is constructed by weighting the number of farm operators in each of seven schooling completion classes: 0–4, 5–7, 8, 9–11, 12, 13–15, $\geqslant 16$ (see U.S. Bureau of the Census, 1967) by years of schooling completed. For a given county, the average education level of black farm operators was assumed to be proportional to the average number of years of schooling completed by all black males 25 years of age and older in 1970 in the county (see U.S. Bureau of the Census, 1972, Table 125).[17] This average schooling level was rescaled so that for each state the derived average education level of black farmers is equal to the state average education level of nonwhite farm operators in the *Census of Agriculture, 1964*.

Extension variables are derived from unpublished federal Extension Service data (see USDA, 1961) on annual time allocations of black and white extension personnel.[18] The simple aggregate extension variable was derived as the annual days devoted to crops, livestock, and planning and management of farm businesses by white and black agents doing primarily agricultural work.

[14] Using sales as the measure of output might reduce the size of blacks' farm output relative to whites' farm output. The average number of persons per household is larger for black operators (4.7) than for white operators (3.4). Thus, black families might be expected to consume a larger share of their farm output. Experiments with output measured as sales and as sales plus home consumption, obtained by distributing the USDA's state level estimates of home consumption among counties on the basis of the number of persons in farm households, showed very similar production function estimates.

[15] With this measure of farmland it was impossible to obtain separate measures of other farmland by race of operator. Land defined to include all land in farms always performed poorly as an input. Its estimated coefficient was unstable in sign and not significantly different from zero.

[16] Although in general black operators have smaller farms than white operators, they also work fewer days per year at nonfarm jobs than white operators (an average of 49.1 for blacks compared with 80.2 for whites). The assumption is that these two differences have approximately offsetting effects on days of farm work.

[17] The share of black rural farm males in all black males is not constant across counties, and education levels differ between farm and nonfarm resident blacks. However, the derived variable seems likely to meet the requirement for an instrumental variable.

[18] Extension data for 1960 (rather than 1963 or 1964) were used because of data availability considerations. One can expect a lag between expenditure of agents' time and the observed effect on agricultural production. Alternatively, one can view the 1960 extension variable as an instrumental variable for extension in a later year, and lagged extension reduces the potential for simultaneous equation bias.

TABLE 2—MEAN VALUE OF INPUTS AND OUTPUT PER FARM: WHITE AND BLACK OPERATOR FARMS IN NORTH CAROLINA, SOUTH CAROLINA, MISSISSIPPI, AND ALABAMA, 1964

Variables	Unit	White Operator Farms	Black Operator Farms
Output (Y)	\$/yr	8,621.3	2,897.4
Machinery ($MACH$)	\$/yr	1,402.8	479.7
Livestock and Feed ($LIVST$)	\$/yr	2,102.9	117.6
Fertilizer and Seed (FS)	\$/yr	622.5	257.6
Farmland[a] (FA)	Acres/yr	75.9	25.3
Operator and Hired Labor (LAB)	Days/yr	314.7	250.2
Schooling (ED)	Yrs.	8.70	5.56
Extension (EXT)	0.1 Days/yr	0.105	0.064
Share Livestock Products in Output (ρ)		0.358	0.058
Share Cropland in Farmland ($A_{.1}/FA$)		0.592	0.740
Share Operator Labor in Operator and Hired Labor (L/LAB)		0.616	0.909

[a] The average number of acres of all land in farms is 168.1 acres for white operators and 49.6 acres for black operators.

Machinery services are measured as the rental on an inventory of a selected group of machines on farms in 1964, plus expenditures on petroleum products and on machinery hire.[19] The livestock and feed input is measured as the rental on the inventory of breeding stock, plus expenditures on purchased livestock and feed. Fertilizer and seed are lumped together and measured as the price-weighted primary plant nutrients, plus the expenditure on seeds. The geographical dummy variables, representing groups of counties with similar soil types, weather in 1964, and general climatic conditions, are state parts of agricultural subregions (see Donald Ibach and James Adams).

Table 2 presents average values for farm output and inputs for white and for black operator farms in North Carolina, South Carolina, Mississippi, and Alabama. These averages by race are obtained by applying the preceding definitions of inputs and output to the state level tables that summarize the characteristics of white and nonwhite operator farms for these states (see U.S. Bureau of the Census, 1967, Tables 18 and 18a). The sample mean values show that black operator farms are on average about one-third as large as white operator farms. Black operator farms produce almost exclusively crops, but livestock products are 36 percent of the farm output of white operator farms. Although black operators have one-third as much farmland per farm as white operators, black operators have a larger (fifteen percentage points) share of their land in cropland. The average schooling level of black operators is 3.14 years lower than for white operators, and the black extension variable is 64 percent as large as the white extension variable. Thus, the average values of inputs and output of black and white operator farms show large differences.

B. *The Estimated Production Function*

The results from fitting the aggregate production function by the method of least squares to the 295 observations are reported in Table 3. The production function was fitted to average per farm values of the levels of the inputs, except for extension which is the county total.[20] The total, rather than average per farm, is relevant if there are large economics of numbers in extending extension information to farmers, for example, by using meetings, demonstra-

[19] Separate machinery data for black and white operator farms do not exist at the county level.

[20] For aggregate data, averages per farm reduce the problem of heteroscedasticity of the random disturbance term in the production function.

TABLE 3—ESTIMATED PRODUCTION FUNCTION FOR SOUTHERN AGRICULTURE: INPUT PRODUCTIVITY DIFFERENCES ON BLACK AND WHITE OPERATOR FARMS, 1964 (295 OBSERVATIONS)

Variables[a]	Coefficients[b]		
	Being Estimated	Estimate	t-ratio
Machinery (*ln MACH*)	α_1	0.313	5.42
Livestock and Feed (*ln LIVST*)	α_2	–	–
Fertilizer and Seed (*ln FS*)	α_3	0.207	4.45
Farmland (*ln FA*)	α_4	0.090	2.49
Share of Cropland in all Farmland $((A_{11}+A_{21})/FA)$	$\alpha_4\left[\frac{a_{11}-a_{12}}{a_{11}^0}\right]$	0.331	3.09
Operator and Hired Labor (*ln LAB*)	α_5	0.614	11.63
Education (*ln ED*)	α_6	2.039	3.11
Share of Black Operators' Schooling in Total Operator Schooling (E_2/ED)	$\alpha_6\left[\frac{e_2-e_1}{e_1^0}\right]$	−0.011	−0.05
Extension (*ln EXT*)	α_7	0.751	3.07
Share of Black Extension in Total Extension (X_2/EXT)	$\alpha_7\left[\frac{f_2-f_1}{f_1^0}\right]$	−0.126	−2.91
(Education) × (Extension) $((ln\ ED) \times (ln\ EXT))$	γ_1	−0.338	−2.89
$\rho \times ln\ LIVST$	β_2	0.620	20.70
$\rho \times ln\ FS$	β_3	−0.253	−3.87
Share of Livestock Products in Farm Output (ρ)	γ_2	−2.890	−5.79
Share of Blacks' Farms in All Farms	γ_3	0.235	1.24
R^2		0.978	
s^2		0.0098	

[a] Output and inputs are county averages per farm, except for *EXT* which is a county total.

[b] Coefficients were estimated for twenty-seven geographical dummy variables. Estimates of these coefficients are reported in Table 4.

tions, and media sources to reach many farmers simultaneously, as opposed to one-to-one consulting. The production function was fitted with an interaction term between education and extension and with two variables that are to capture residual effects of product mix (ρ) and racial mix of farm operators on farm output.

Several specifications of the basic equation were fitted to check on consistency of estimated coefficients across variables. In the final regression, consistency across estimated coefficients is imposed in the sense that, if the direct estimate of β_i (or of a relative productivity coefficient) was not significantly different from zero, then coefficients to be estimated that contained β_i as one part of a product were set equal to zero.

The results show that parameters of the production function differ by product mix of output. As the share of livestock output in total farm output increases, the coefficient of the livestock (fertilizer and seed) input increases (decreases). When the livestock output is zero and crop output is positive, the coefficient of the livestock input is zero; clearly a plausible finding. The coefficient of the fertilizer and seed input is largest when only crop output is produced, and it declines as the share of livestock output increases. The decline is plausible because livestock manures can substitute for commercial fertilizer in crop production. The negative and significant coefficient of ρ implies that the (constant of the) production function shifts down as the share of livestock products in total output increases. This effect seems to reflect the greater use of inputs for maintenance in livestock than in crop production.[21]

The estimated coefficients of the education and extension variables are all significantly different from zero, except for the extension ratio term. The estimated coefficient of the education-extension interaction effect is negative, suggesting that farmers' education and agricultural extension are substitutes in southern agricultural production in the sense that higher education (extension) levels reduce the coefficient of extension (education). At sample mean values, the estimated coefficient of education is $0.058\ [=2.038-0.338(5.861)]$ and of extension is $0.051\ [=0.751-0.338(2.071)]$.

The coefficients of black operators' schooling share and of black extensions' share of total extension are estimates of the average relative quality differences of a unit of black compared with a unit of white schooling and extension, respectively. Given an estimated education coefficient of 0.058 at the sample mean, the estimated coefficient of black operators' schooling share of -0.011 implies that the average quality of a year of black operators' schooling as it affects agricultural production is 19 percent lower than white operators' schooling. The estimated coefficient is, however, not significantly different from zero. Thus, for effects on agricultural productivity, the primary black-white schooling difference is from years of schooling completed and not from schooling quality.

The coefficient of black extensions' share of total extension days is negative and significantly different from zero at the 5 percent level. Given the estimated extension coefficient of 0.051 at the sample mean, the estimated coefficient of the black extension share of -0.126 implies that the average quality of a day of black extension as it affects agricultural production is 247 percent lower than a day of white extension. Thus for effects on agricultural productivity, both low quality and quantity of black extension input are sources of black-white differences.[22]

Other results are that cropland is significantly more productive than other (nonwoodland pasture) land. Given the estimated coefficient for land of 0.090, the estimated coefficient of the share of cropland in all farmland of 0.331 implies that an acre of cropland is 3.68 times more productive than an acre of nonwoodland pasture. Also, there is no significant difference in the productivity of a day of operator labor compared with hired labor, of a day of black operator labor compared with white operator labor, or of an acre of black operator cropland compared with white operator cropland. The positive but not significantly different from zero coefficient of the variable "share of farm operators that are black" suggests that major black-white productivity differences have been accounted for by other included variables.[23]

IV. Implications and Conclusions

It is well known that blacks have been discriminated against historically in quantity and quality of educational opportunities. This study has illuminated the dis-

[21] The economic significance of the nonzero βs is that they imply optimal relative input combinations change when the output mix changes, holding relative input prices constant.

[22] In 47 percent of the sample counties, black farmers were present, but black extension personnel were not. A dummy variable was used to test for an effect on farm output of absence of black extension to assist black farmers. The coefficient of this dummy variable was generally negative but not significantly different from zero. Thus, the effects of black extension on farm output seem to be adequately represented by the ratio of black to total extension input.

[23] When the full model was estimated, the coefficients containing β_4 and β_5 (for example, $\beta_4 c_1, \beta_4 c_2, \beta_4 c_3$, etc.) were not significantly different from zero. When the variables associated with these coefficients were excluded from the full model, the estimated coefficients of A_{21}/FA and L_2/LAB were not significantly different from zero. Furthermore, the null hypothesis that the coefficients of the eleven variables excluded from the full model to obtain the reported production function are simultaneously equal to zero cannot be rejected at the 5 percent significance level. The calculated F-value is 0.89, and under standard least squares assumptions, the tabled F-value for 11 and 243 degrees of freedom at the 5 percent level is 1.82.

TABLE 4—ESTIMATES OF THE COEFFICIENTS OF THE GEOGRAPHICAL DUMMY VARIABLES IN THE PRODUCTION FUNCTION

Variables	Number of Counties in Subregion	Percent Nonwhite Farms in Subregion	Percent of Sample Nonwhite Farms in Subregion	Coefficients	
				Estimate	*t*-ratio
South Carolina					
SASR 15[a]	7	52.4	3.0	−3.605	−2.64
SASR 16	5	40.4	5.1	−3.452	−2.51
SASR 26	12	15.5	2.0	−3.565	−2.62
SASR 27	5	25.1	1.2	−3.550	−2.59
SASR 28	11	49.7	5.6	−3.544	−2.58
SASR 33	6	26.4	1.1	−3.582	−2.62
North Carolina					
SASR 14	9	46.4	4.3	−3.267	−2.38
SASR 15	18	25.3	4.5	−3.251	−2.37
SASR 16	6	42.9	3.6	−3.192	−2.33
SASR 17	14	27.7	8.9	−3.189	−2.33
SASR 18	3	19.8	4.6	−3.073	−2.25
SASR 25	22	0.8	0.2	−3.334	−2.46
SASR 26	18	6.8	1.5	−3.374	−2.47
Mississippi					
SASR 31	10	5.7	0.4	−3.580	−2.63
SASR 46	9	12.9	1.6	−3.303	−2.42
SASR 47	7	43.5	4.2	−3.449	−2.53
SASR 48	34	30.9	14.1	−3.461	−2.55
SASR 49	11	55.6	8.8	−3.351	−2.44
SASR 64	11	55.4	5.7	−3.347	−2.44
Alabama					
SASR 31	3	8.0	0.3	−3.443	−2.52
SASR 32	12	21.0	3.0	−3.521	−2.57
SASR 33	8	23.7	1.5	−3.431	−2.52
SASR 34	8	7.0	0.6	−3.440	−2.52
SASR 45	11	5.8	1.4	−3.335	−2.44
SASR 46	12	20.2	2.7	−3.436	−2.52
SASR 47	10	66.2	8.9	−3.567	−2.61
SASR 48	3	36.8	1.1	−3.567	−2.63
Total	295	26.6	100.0		

[a] *SASR* = State part of an agricultural subregion (see Ibach and Adams).

crimination against southern black farmers in quantity and quality of public agricultural extension assistance provided to them. The results from the estimated production function provide empirical support for the hypothesis of lower relative productivity or quality of black farmers' schooling and black extension compared with white farmers schooling and white extension.

The estimated production function (Table 3) and available state level data permit a comparison of productivity differences for black and white operator farms. Marginal products for black and for white operator farms are evaluated at their respective sample means for inputs (Table 2).[24] The implied marginal product of labor on black operator farms is only 42 percent as large as for white operator farms. The size of this black-white difference is consistent with urban black-white wage differences of this period (see Welch, 1973a). Although the estimated production function showed that quality per unit of black farmers' education and black extension is lower than for whites,

[24] The constant term from Table 4 is −3.390 for blacks and −3.370 for whites. These were obtained by weighting coefficients of *SASR*s by the actual distribution of farms by race.

the implied marginal products of black operators' schooling and of black extension are about three and ten times as large as the marginal product of white operators' schooling and extension, respectively.[25] The reasons for these large differences are the relatively small size of average black education and extension inputs, given diminishing marginal productivity, and the negative education-extension interaction effect in production. These positive effects more than offset the negative effects on output of lower quality of black schooling and extension. The positive difference in black-white marginal product of education is in direct contrast to educations' effect on rural income differences (see Welch, 1967). The implied marginal products for other inputs are essentially equal across the two races.

We can estimate the effect on total factor productivity if all farms were suddenly operated by blacks. In making this comparison, I attribute all productivity differences to effects of differences in quantity and quality of schooling and extension, and to racial mix of operators. Other changes in the switch are assumed to have neutral effects on the total productivity differential. The contribution of lower average schooling levels of black operators relative to white operators and of lower average black extension input relative to white is -26.7 percent [$= 2.039\,(-0.447) + 0.751\,(-1.342) - 0.338\times(-4.883)$]. The contribution of lower average quality per year of black schooling and per day of black extension is -13.7 percent ($= -0.011 - 0.126$). Because all farms would now have black operators, the coefficient of the share of farmers that are black, γ_3, contributes 23.5 percent to the black-white farm productivity differential. Thus, the evidence is that if all white farms were suddenly operated by blacks, southern farm output would be about 17 percent lower than if all farms were operated by whites. The lower level of schooling and extension for black farmers than for white farmers would contribute twice as much to this productivity differential as lower quality of schooling and extension.

[25] These ratios were calculated as follows: for education as $[(2.039-0.338\times4.428)\hat{Y}_B/5.56]/[(2.039-0.338\times5.770)\hat{Y}_W/8.70]$ and for extension as $[(0.751-0.338\times1.716)\hat{Y}_B/83.76]/[(0.751-0.338\times2.163)\hat{Y}_W/320.67]$ where $\hat{Y}_W$ and $\hat{Y}_B$ are the imputed values of output for white and black operator farms, respectively.

REFERENCES

R. Day, "The Economics of Technological Change and the Demise of the Sharecropper," *Amer. Econ. Rev.*, June 1967, *52*, 427–49.

Stephen J. DeCanio, *Agriculture in the Postbellum South: The Economics of Production and Supply,* Cambridge 1974.

Edward D. Eddy, *Colleges For Our Land and Time: The Land-Grant Idea in American Education*, New York 1957.

R. E. Evenson, "Economic Aspects of the Organization of Agricultural Research," in Walter L. Fishel, ed., *Resource Allocation in Agricultural Research,* Minneapolis 1971.

______, "A Century of Agricultural Research and Productivity Change Research, Invention, Extension and Productivity Change in U.S. Agriculture: An Historical Decomposition Analysis," in Ahmed Araji, ed., *Research and Extension Productivity in Agriculture*, Moscow, Idaho 1980, 146–228.

F. M. Fisher, "The Existence of Aggregate Production Functions," *Econometrica*, Oct. 1969, *37*, 553–77.

Robert R. Fogel and Stanley Engerman, *Time on the Cross: The Economics of American Negro Slavery*, Boston 1974.

John H. A. Green, *Aggregation in Economic Analysis: An Introductory Survey,* Princeton 1964.

Zvi Griliches, "Estimates of the Aggregate Agricultural Production Function From Cross-sectional Data," *J. Farm Econ.*, May 1963, *45*, 419–32.

______, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, Oct. 1957, *25*, 501–22.

______, "Research Expenditures, Education and the Agricultural Production Function," *Amer. Econ. Rev.*, Dec. 1964, *54*, 961–74.

______ **and Vidar Ringstad**, *Economies of Scale and the Form of the Production Function,* Amsterdam 1971.

R. E. Hall, "The Specification of Technology with Several Kinds of Output," *J. Polit. Econ.*, July/Aug. 1973, *81*, 878–92.

I. Hoch, "Simultaneous Equation Bias in the Context of the Cobb-Douglas Production Function," *Econometrica*, Oct. 1958, *26*, 566–78.

W. E. Huffman, "Allocative Efficiency: The Role of Human Capital," *Quart. J. Econ.*, Feb. 1977, *91*, 59–79.

______, "The Productive Value of Human Time in U.S. Agriculture," *Amer. J. Agri. Econ.*, Nov. 1976, *58*, 672–83.

Donald B. Ibach and James R. Adams, *Crop Yield Response to Fertilizer in the United States*, Statist. Bull. 431, U.S. Dept. Agriculture, Washington, Aug. 1968.

J. D. Reid, "Sharecropping and Agricultural Uncertainty," *Econ. Develop. Cult. Change*, Apr. 1976, *24*, 549–76.

______, "The Theory of Share Tenancy Revisited—Again," *J. Polit. Econ.*, Apr. 1977, *85*, 403–08.

Walter E. Sellers, *Variation in Length of Farm Workweek*, Statist. Bull. 474, U.S. Dept. Agriculture, Washington, Sept. 1971.

Finis Welch, (1973a) "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *63*, 893–907.

______, (1973b) "Education and Racial Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets,* Princeton 1973.

______, "Education in Production," *J. Polit. Econ.*, Jan./Feb. 1970, *78*, 59–79.

______, "Labor-Market Discrimination: An Interpretation of Income Differences in the Rural South," *J. Polit. Econ.*, June 1967, *75*, 225–40.

A. Zellner, J. Kmenta, and J. Drèze, "Specification and Estimation of Cobb-Douglas Production Function Models," *Econometrica*, Oct. 1966, *34*, 784–95.

U.S. Bureau of the Census, *Census of Agriculture, 1964*, Vol. I, Washington 1967; Vol. II, 1968.

______, *U.S. Census of Population: 1970*, Washington 1972.

U.S. Commission on Civil Rights, *Equal Opportunity in Farm Programs*, Washington 1965.

______, *The Federal Civil Rights Enforcement Effort—A Reassessment,* Washington, Jan. 1973.

U.S. Department of Agriculture, "Annual Report by County Extension Agents," unpublished records, Washington 1961.

______, *Farm Labor*, Washington, Jan. 1965.

U.S. Office of Education, *Education of Negro Leaders,* Bull. 1948, No. 3, Washington 1949.

______, *Statistics of Land-Grant Colleges and Universities*, circular no. 689, Washington 1961.

______, *Statistics on Land-Grant Colleges and Universities*, Bull. 1, 4, 10; Washington 1947, 1951, 1956.

Control and Decontrol of Wages in the United States: An Empirical Analysis

By FRANK REID*

An overall assessment of a controls program requires consideration of a wide range of potential costs and benefits. This paper is directed toward the much narrower objective of measuring one aspect of the potential benefits of controls, namely the reduction in the rate of change of wages (and prices) below what they otherwise would be, both during and after controls. Two episodes of controls are considered: the guidepost policy of the 1960's and the four phases of controls imposed by the Nixon Administration during the period August 1971 to April 1974.

It is clear that one cannot determine the effect of controls on the rate of wage change by simply comparing the rate of change during controls to the rate of change immediately preceding controls. The reason is that other factors which affect the rate of wage change may have varied—for example, the labor market may have been affected by a change in government monetary or fiscal policy. In order to assess the effect of controls, an economic model is required to predict the rate of wage change which would have occurred in the absence of controls.

Similarly the effect of controls cannot be judged by comparing the actual rate of change during controls with the target rate of change or "guideline" established by the government. The reason is that, if, on the one hand, the target is achieved, it may be due to basic economic factors independent of any contribution of controls. If, on the other hand, the guideline is exceeded, it is still possible that the rate of wage change is lower than the rate that would have occurred in the absence of controls, and thus controls may have had some restraining influence.

An argument frequently encountered is that if controls are effective in restraining the rate of wage and price change while in force, their removal will be followed by an "explosion" in the rates of wage and price change that will partially or totally offset any effects of controls while in force. One of the novel aspects of the present paper is that this explosion hypothesis is statistically tested by measuring both the reduction (if any) in the rate of wage change while controls are in force, and any subsequent increase in the rate of change when controls are removed.

The model used to predict the rate of wage change that would have occurred in the absence of controls is specified in Section I. In Section II the effect of controls and any subsequent explosions are estimated using the conventional techniques of examining the residuals from the controls-off equation and of including intercept shift dummies to model the effects of controls. The results are then compared with results from previous studies.

Some criticisms of the intercept-dummy method are offered in Section III and a more sophisticated method of modelling the effects of controls is suggested. The method is based on the "rotation hypothesis" which postulates that the amount of restraint due to controls is a linear function of the difference between the guideline and the rate of increase predicted in the absence of controls. It is a more general technique which includes the intercept shift dummy method as a special case. Section IV contains conclusions and policy implications.

I. The Controls-Off Wage Equation

In order to predict the rate of wage change that would have occurred in the absence

*University of Toronto. I am indebted to Dick Lipsey, Jim Pesando, Schmuel Sharir, David Smith, and Philip Smith for helpful comments. The paper has also benefited from the comments of George Borts and an anonymous referee. Any remaining defects are, of course, my own responsibility.

of controls (W_t^n), the full sample period (1960:I–1978:I) is divided into controls-off and controls-on subperiods, and a wage equation is estimated for the controls-off sample. The controls-off sample excludes the guideposts period (1962:I–1966:IV), the Nixon controls period (1971:IV–1974:I), and any quarters following controls in which an explosion occurred. Some preliminary analysis indicated that the three quarters following the termination of Phase IV should be excluded on these grounds (1974:II–1974:IV), but there was no indication of any significant effect following the termination of the guideposts policy. Thus the controls-off sample consists of forty observations (1960:I–1961:IV, 1967:I–1971:III, and 1975:I–1978:I), and the controls-on sample (including explosions) consists of the remaining thirty-three observations. All data used are seasonally adjusted. For details of data sources and definitions see the Appendix.

The dependent variable is the quarter-to-quarter percentage change at an annual rate in the index of hourly earnings in manufacturing, adjusted for overtime and interdustry shifts.[1] Economic theory suggests that the expected inflation rate (PE_t) should enter as an explanatory variable with a coefficient of unity (see Edmund Phelps and Milton Friedman). In this paper two alternative proxies for expected inflation are employed, the Livingston survey data on expected inflation (as reworked by John Carlson), and an expected-inflation series constructed according to a weak form of the rational expectations hypothesis. In constructing this series, expectations were assumed to be "efficient" in the sense that predictions make full use of the past history of inflation.[2] More specifically, an expected-inflation variable was calculated as the predicted value from a regression of the current quarter-to-quarter percentage change in the Consumer Price Index (*CPI*) on inflation rates in the previous four quarters, with weights constrained to follow a linear pattern.[3]

Empirically, the two expected-inflation series are highly correlated ($r=0.93$) and give very similar results concerning the effects of controls. Results using rational expectations are reported in the body of the paper and key results using the reworked Livingston series are reported in a footnote (since the latter are available only to 1976:I). An important advantage of both of these expected-inflation rate series is that since the current inflation rate does not appear as an explanatory variable in the wage equation, the usual wage-price simultaneity problem is avoided.

Theoretical work suggests both the level and rate of change of excess demand for labor as explanatory variables in the wage equation (see, for example, A. W. Phillips and Phelps). As a proxy for excess demand in the labor market, both the unemployment rate and the vacancy rate were tried. It was found that the vacancy rate performed somewhat better in terms of explanatory power and statistical significance. The vacancy rate series was constructed using the seasonally adjusted monthly index of Help-Wanted Advertising published in *The Conference Board Statistical Bulletin*. A proxy for the monthly job vacancy rate was constructed as $V_t=(x_t L_{1967})/L_t$, where x_t is the Help-Wanted Index in month t (base

[1] Throughout this paper percentage changes are calculated as

$$y_t=100((1+(x_t-x_{t-1})/x_{t-1})^4-1)$$

where x_t is the level during quarter t and y_t is the percent change.

[2] Such an autoregressive expectations function will be rational in the sense of John Muth under very restrictive assumptions. It is, however, an optimal forecasting function in a statistical sense for a wide range of stochastic processes. The rational expectations series was constructed using a one-quarter horizon since previous empirical work with wage equations has indicated that there is little to be gained by using a "chaining principle" to construct an expected-inflation series with a horizon closer to the average contract length.

[3] Lags varying from zero to twelve quarters were scanned, and distributed lag weights were tried both unrestricted and constrained to lie on polynomials of the first, second, and third degree. Maximum explanatory power was achieved with a four-quarter lag. The *F*-tests indicated that at the .05 significance level the linear restriction on the weights could not be rejected, and higher-order polynomials did not add significantly to explanatory power.

1967=100), L_t is the civilian labor force in month t, and L_{1967} is the annual labor force in 1967. Two functional forms were tried, the vacancy rate and its inverse, $VIN_t = 100/V_t$. The latter was found to have greater explanatory power. Distributed lags on both VIN_t and its first difference, $DVIN_t$, were tried varying in length from zero to twelve quarters and with weights both unconstrained and constrained to lie along polynomials of first, second, and third degree. The most explanatory power was obtained using only the current and one-quarter lag value of VIN_t.

The use of the vacancy rate avoids several criticisms which have been made of the use of the unemployment rate in wage equations as an indicator of labor market conditions. Norman Simler and Alfred Tella argue that the meaning of the unemployment rate has been affected by trends in participation rates which have changed the amount of hidden unemployment. George Perry constructs an adjusted unemployment measure in which the unemployment rates of various demographic groups are weighted to reflect differences in average number of hours worked and hourly earnings. More recently, Walter Oi has noted the changed meaning of the unemployment rate due to the extended coverage of unemployment insurance programs and the increase in the ratio of benefits to average weekly earnings.

To summarize, the theoretical specification of the controls-off wage equation is

$$(1) \qquad W_t^n = \beta_0 + \beta_1 PE_t + \beta_2 VIN_t + \beta_3 VIN_{t-1} + \varepsilon_{1t}$$

where ε_{1t} is an error term with the usual properties and the a priori expectations regarding the coefficients are $\beta_1 = 1, \beta_2 + \beta_3 < 0$. The form of this equation is strikingly simple—no variables of dubious theoretical validity have been inserted in an *ad hoc* manner. This is in accord with Oi's suggestion that more emphasis be placed on theoretical rather than statistical considerations in determining which variables should enter the wage equation.

Empirical estimation of (1) for the controls-off sample yields

$$(2) \quad W_t^n = \underset{(4.3)}{3.84} + \underset{(9.5)}{0.84}\, PE_t - \underset{(-2.9)}{5.65}\, VIN_t + \underset{(2.5)}{4.98}\, VIN_{t-1} \qquad R^2 = 0.76 \quad d = 1.97$$

where the figures in parentheses are t-statistics and d is the Durbin-Watson statistic.

The expected inflation variable PE_t has a coefficient of 0.84 and is highly significant.[4] A (two-tailed) t-test indicates that the coefficient of PE does not differ significantly from its theoretically expected value of unity ($t = 1.79$). Both VIN_t and VIN_{t-1} are significant and VIN_t has a negative coefficient greater in absolute value than the positive coefficient on VIN_{t-1}. This suggests that both the level of vacancies and the rate of change of vacancies are important determinants of wage increases.

As a test for structural stability, the controls-off sample was divided into two equal subperiods (1960:I–1969:IV and 1970:I–1978:I) and reestimated. The null hypothesis of structural stability could not be rejected at the .05 significance level.

II. Modelling the Effects of Controls using Intercept Shift Dummies

In this section, the effects of controls are determined by examining the residuals during various phases of controls, and an attempt is made to model the observed effects using intercept shift dummies.

Information on the residuals during the various periods of controls and subsequent explosions is given in Table 1. The results indicate that during the guideposts period the rate of wage change was, on average, 1.5 percentage points below the predicted rate of change. In the year following the termination of the guideposts the actual rate of wage change equalled the predicted rate of

[4] Unless otherwise indicated, hypotheses are tested using a .05 significance level, and either a one-tail or two-tailed t-test as appropriate.

TABLE 1—ACTUAL AND PREDICTED MEAN RATE OF WAGE CHANGE DURING VARIOUS CONTROLS PERIODS

Period	Symbol	Date of Period	Actual Rate of Wage Change	Predicted from Controls-Off Equation	Residual
Guideposts	*D*0	1962:I–1966:IV	2.8	4.3	−1.5
Guideposts Explosion	*E*0	1967:I–1967IV	4.9	4.9	0.0
Phase I	*D*1	1971:IV	3.0	6.7	−3.7
Phase I Explosion	*E*1	1972:I	9.9	5.0	+4.9
Phase II	*D*2	1972:II–1972:IV	5.7	5.7	0.0
Phase III	*D*3	1973:I–1973:II	5.9	6.3	−0.4
Phase IV	*D*4	1973:III–1974:I	7.2	9.2	−2.0
Phase IV Explosion	*E*4	1974II–1974IV	11.5	10.6	+0.9

change (4.9 percent) giving no indication of an explosion following the guideposts period.[5]

For the Phase I freeze, the rate of wage change predicted from the controls-off equation is 6.7 percent and the actual rate of change was only 3.0 percent indicating a substantial reduction during this controls period. In the quarter following Phase I, however, the wage increases which were deferred in Phase I took effect in addition to the normal round of new settlements resulting in an explosion in the rate of wage change following Phase I. Table 1 indicates that the magnitude of this explosion (4.9 percentage points) is such that it completely offset the reduction during Phase I.

During Phase II, a guideline for wage increases of 5.5 percent was established and administered by the Pay Board. This guideline was retained during the switch to the voluntary controls of Phase III and the return to compulsory controls in Phase IV. The data in Table 1 indicate that the mean predicted rate of wage increase during Phase II was 5.7 percent, only slightly above the guideline. The actual rate of wage change was also 5.7 percent, giving no indication of a restraining effect on wages during Phase II.

The mean predicted rate of wage change for Phase III is 6.3 percent. Comparison with the actual mean rate of wage change of 5.9 percent indicates a small restraining effect of about 0.4 percentage points.

The actual mean rate of wage change during Phase IV was 7.2 percent, substantially above the guideline but considerably below the 9.2 percent increase predicted for Phase IV. Thus, on average, Phase IV restrained wage increases by about 2.0 percentage points. In the three quarters following Phase IV, however, the mean rate of wage change was 0.9 percentage points greater than predicted, partially offsetting the restraint during Phase IV.

Compared to other methods of estimating the effects of controls, the examination of mean residuals has the advantage of being the most straightforward technique. The use of intercept shift dummies, however, has a number of advantages over the simple examination of residuals: (i) It allows a convenient test of the statistical significance of any observed reductions or explosions. (ii) The estimates are slightly more efficient because of the additional degrees of freedom obtained by using the controls-on observations to estimate the parameters of the wage equation. (iii) It allows prediction of the future effects of controls, given values of the explanatory variables. The disadvantage of the intercept shift dummy, however, is that it requires the assumption that controls do not affect the slope coefficients of the wage equation. This point is considered further in Section III below.

To explain the rate of wage change over the full sample period (W_t), equation (1)

[5] There is some arbitrariness in deciding the period over which to measure an explosion. Similar results were obtained examining the mean for two years following controls.

was estimated including shift dummies for the various periods of controls and explosions indicated in Table 1. The result is

$$(3)\quad W_t = \underset{(4.4)}{3.61} + \underset{(10.3)}{0.84}\,PE_t - \underset{(-3.4)}{4.96}\,VIN_t$$

$$+ \underset{(2.9)}{4.48}\,VIN_{t-1} - \underset{(-4.5)}{1.4}\,D0 + \underset{(0.1)}{0.1}\,E0$$

$$- \underset{(-4.1)}{3.7}\,D1 + \underset{(4.7)}{4.9}\,E1 + \underset{(0.2)}{0.1}\,D2$$

$$- \underset{(-0.5)}{0.3}\,D3 - \underset{(-3.3)}{1.9}\,D4 + \underset{(1.3)}{0.9}\,E4$$

$$R^2 = 0.89 \quad d = 2.07$$

The overall explanatory power of the equation is high, there is no evidence of significant autocorrelation, and the estimated coefficients of the economic variables are similar to those obtained for the controls-off period.

The coefficients of the shift dummies in equation (3) indicate effects which are very close (within one-tenth of a percentage point) to those indicated in Table 1. One-tailed t-tests indicate that the observed reductions during the guideposts, Phase I, and Phase IV are significant. During Phase II and Phase III, there was no significant effect on the rate of wage change.

The coefficient of $E1$ indicates a statistically significant explosion following Phase I and a t-test indicates that one cannot reject the null hypothesis that the coefficients of $D1$ and $E1$ sum to zero ($t = 0.88 < t^* = 2.0$). The coefficient of $E4$ indicates that the explosion following Phase IV was statistically significant at the .10 level, but not the .05 level (one-tailed t-test). The null hypothesis that the coefficients of $D4$ and $E4$ sum to zero could not be rejected at the .05 significance level ($t = 1.06 < t^* = 2.0$). Thus the evidence is consistent with the view that wage explosions following Phase I and Phase IV completely offset any significant restraining effect which the Nixon controls had while in force.[6]

The effects of controls estimated above include both the direct effect of controls on the rate of wage change (given the values of the explanatory variables) and any indirect effects arising from an announcement effect of controls on expected inflation.

Although the available empirical evidence regarding an announcement effect of controls on expected inflation is not conclusive, most studies indicate that any effect is either small or nonexistent. For example, Carlson and Michael Parkin conclude from their analysis of inflationary expectations data obtained from the Gallup Poll survey in the United Kingdom that five periods of wage and price controls in the 1961–1973 period "...appear to have trivial and totally insignificant effects on expectations" (p. 133). They found that expected inflation depends only on past actual inflation with no announcement effect due to controls. Similarly, George De Menil and Surjit Bhalla analyze survey data on expected price changes from the *Survey of Consumer Finance* and conclude that "Pronouncements designed to bring down inflationary expectations—of which there have been many in the United States recently—may have little lasting effect if they do not also reduce actual inflation" (p. 178).

Timothy McGuire uses three alternative methods to estimate the direct announcement effect on expectations: on the basis of interest rate behavior, expected inflation appears to have been reduced by about 0.8 percentage points during the first half of the Nixon controls and to have been increased by about 1.0 percentage points in the second half. Analysis using the Livingston data showed a reduction of about 1.5 percentage

[6] The cumulative reduction in wage inflation due to controls equals $0.25 d_j x_j$, where d_j is the coefficient of D_j and x_j is the length of the period of controls in quarters. Similarly, the cumulative effect of the explosion is given by $0.25 e_j y_j$ where e_j is the coefficient of E_j and y_j is the length of the explosion. For Phase I and Phase IV it happens that $x_j = y_j$, implying that the net effect of the controls and explosion in each of these two periods can be calculated as 0.25 times the sum of the coefficients of D_j and E_j.

points in the first half of the Nixon controls and a roughly equivalent increase in the second half. Using a comparison of deterministic (*ex ante*) expected inflation forecasts with conditional (maximum likelihood) forecasts, McGuire found an insignificant reduction of about 0.4 percentage points in expected inflation in 1971:4 and no effect subsequently. This evidence suggests that the observed reductions in the rate of wage change due to controls are primarily due to the direct effect of controls, and any indirect effect on expectations is of small empirical magnitude.

The conclusions on the effects of controls presented in this section are broadly consistent with most previous studies which have used residuals or shift dummies to estimate the effects. Studies which report significant negative coefficients on guidepost dummies include Ronald Bodkin et al. (−1.8), Susan Vroman and Wayne Vroman (about −.72 when converted to annual rates), Robert Gordon (−1.8 in his 1972 reestimation of the Eckstein-Brinner model, −0.7 in his reestimation of the Perry model), Oi (−1.8 in his 1976 reestimation of the Perry model). Gordon (1971) and Stanley Black and H. H. Kelejian find insignificant effects.

Vroman and Vroman calculate mean residuals for various periods related to the Nixon controls. Using an average hourly earnings equation, they find insignificant positive residuals for the Phase I freeze and Phase I explosion combined, Phase II, and the Phase IV explosion. They find an insignificant negative residual for Phases III and IV combined. Using an *ARIMA* model, Edgar Feige and Douglas Pearce find a significant reduction of 4.4 percentage points during Phase I and an insignificant increase of 1.2 percentage points for Phase II and the Phase I explosion combined. (Averaging my estimated Phase I explosion of 4.9 percentage points over the four quarters of Phase II results in a mean residual of 1.25 percentage points.) For the combined effects of Phases I, II, and III (and the first quarter of Phase IV), Gordon finds a very small positive mean residual of 0.06 (see Gordon 1973, Table 1, simulation *B*). This is consistent with the results in the present paper that these three phases of the Nixon controls (including the Phase I explosion) resulted in little or no significant reduction in the rate of wage change.

III. Modelling the Effects of Controls Using the Rotation Hypothesis

The use of shift dummies to measure the effect of controls has been criticized on two grounds. From a statistical viewpoint, Oi points out that the validity of the intercept shift dummy technique depends on the assumption that controls do not affect the slope coefficients of the wage equation. Oi suggests that this assumption should be subjected to a statistical test and, as an example, reestimates the wage equation in Perry. He finds that for the guideposts period, an *F*-test rejects the equal slope assumption and thus the dummy variable technique.

Richard Lipsey and Parkin have criticized the use of shift dummies on theoretical grounds. They hypothesize that if the government establishes some guideline for wage increases W_t^*, which is intended as a maximum but is not rigidly enforced, then the amount of restraint will be related to the amount by which wage increases would have exceeded the guideline in the absence of controls. They also hypothesize that there will be no restraint on wage settlements which would have been below the guideline, and that if the guideline acts to some extent as a minimum target in wage bargaining, the controls policy may actually increase settlements in this range.

In graphical terms, since the hypothesis implies that controls reduce settlements above W_t^* and increase (or do not affect) settlements below W_t^*, the effect is clearly to rotate the wage equation at W_t^*. For this reason I shall refer to the Lipsey-Parkin view as the rotation hypothesis. If the rotation hypothesis is correct, it implies that controls will affect both the intercept and the slope coefficients of the wage equation, and thus the use of simple shift dummies is inadequate. In his paper Oi refers to the Lipsey-Parkin argument as a possible theoretical explanation for his empirical finding that controls affect the slope coefficients.

Lipsey and Parkin attempted to test their hypothesis simply by fitting separate wage (and price) equations to the controls-off and controls-on periods. Although they found significant differences between the periods, critics have pointed out substantial econometric problems with their work: serious autocorrelation induced by the use of a four-quarter overlapping dependent variable, simultaneous equation bias, a very poor fit of the controls-on equation, and lack of robustness of the results. (See for example, the excellent survey of the econometric evidence by Parkin, Michael Sumner, and Robert Jones.)

Lipsey and Parkin's basic idea of the rotation of the wage equation, however, has been extended in my earlier work. I developed a simple working hypothesis that the amount of restraint due to controls is proportional to the difference between the guideline and the rate of increase which would have occurred in the absence of controls. That is,

$$(4) \qquad W_t^n - W_t^c = k(W_t^n - W_t^*) + \varepsilon_{4t}$$

where W_t^n is the rate of increase predicted in the absence of controls and W_t^c is the rate of wage change during controls. The proportion of the excess eliminated, k, was called the *effectiveness coefficient* of controls.

More generally, it is hypothesized that the reduction due to controls is given by

$$(5) \qquad W_t^n - W_t^c = -d + k(W_t^n - W_t^*) + \varepsilon_{5t}$$

This linear rotation hypothesis encompasses the Lipsey-Parkin proportional rotation hypothesis and the conventional shift dummy as special cases in which $d=0$ and $k=0$, respectively.

For the wage model used in the present paper, the null hypothesis that the slope coefficients of the wage equation are equal in the controls-off period and the controls-on subperiods (excluding explosions) was tested and rejected at the .05 significance level $\{F=2.29>F^*_{.95}(6,57)=2.25\}$.

To determine if the rotation hypothesis could account for this difference equation (5) was estimated for each of the two controls-on subperiods, using the estimate of W_t^n from the controls-off equation. The results are, for the guideposts,

$$(6) \qquad W_t^n - W_t^c = \underset{(4.8)}{1.2} + \underset{(1.2)}{0.22}\,(W_t^n - W_t^*)$$

$$R^2 = .07$$

and, for the Nixon controls (excluding explosions),

$$(7) \qquad W_t^n - W_t^c = \underset{(-0.4)}{0.1} + \underset{(6.6)}{0.57}\,(W_t^n - W_t^*)$$

$$R^2 = 0.86$$

These results indicate that for the guideposts the rotation hypothesis is not supported—there is no significant relationship between the amount of the reduction and the amount of the excess, $W_t^n - W_t^*$. For the Nixon controls, however, the proportional rotation hypothesis receives strong support: The estimate of k is highly significant and indicates that the Nixon controls eliminated about half of the difference between W_t^n and the guideline. The F-tests also clearly indicated that there was no significant difference between the four phases of the Nixon controls in the estimates of equation (7).

The implications of the linear rotation hypothesis for modelling the rate of wage change during controls will now be analyzed. In general, consider a controls-off wage equation in which W_t^n is a linear function of n variables (including the constant):

$$(8) \qquad W_t^n = \sum_{1=1}^{n} \beta_i X_{it} + \varepsilon_{8t}$$

Substituting (8) into (5) and solving for W_t^c yields

$$(9) \qquad W_t^c = kW_t^* + \sum_{i=1}^{n} (1-k)\beta_i X_{it} + d + \varepsilon_{9t}$$

where $\varepsilon_{9t} = (1-\mathrm{k})\varepsilon_{8t} + \varepsilon_{5t}$

The linear rotation hypothesis, if valid, has two important implications for the specification of the wage equation during controls: (i) In addition to the set of variables which appear in the controls-off equation, an intercept shift dummy and the wage guideline W_t^* should appear as an independent "variable" in the equation. (ii) The theoretical analysis implies certain restrictions on the estimated regression coefficients in the controls-on equation. The restriction is that the coefficients of the explanatory variables should equal $(1-k)$ times the coefficients in the controls-off equation, where k is the coefficient of W_t^*.

Generalizing to m periods of controls and taking account of the possibility of explosions, the implied theoretical constraints can be imposed by estimating the following equation for the full sample period, where D_{jt} is a set of dummy variables which take the value unity during controls period j and zero otherwise, and E_{jt} is a set of dummies which take the value unity during explosion period j and zero otherwise ($j=1,\dots,m$):

$$(10)\quad W_t=\sum_{i=1}^{n}\beta_i X_{it}+\sum_{j=1}^{m}d_j D_{jt}+\sum_{j=1}^{m}e_j E_{jt}$$

$$+\sum_{j=1}^{m}k_j D_{jt}W_{jt}^*-\sum_{j=1}^{m}\sum_{i=1}^{n}k_j\beta_i D_{jt}X_{it}+\varepsilon_{10,t}$$

It is clear that estimation of (10) requires a set of non-linear constraints on the estimated parameters. Although equation (10) contains $n+2m+mn$ variables, there is not a serious reduction in degrees of freedom because, due to these theoretical constraints, the number of parameters estimated is only $n+3m$.

The interpretation of (10) can be facilitated by considering some special cases. If controls are totally ineffective, i.e., $d_j=k_j=e_j=0$ for all j, then (10) collapses to (8), the controls-off wage equation. If controls do not rotate the wage equation, i.e., $k_j=0$ for all j, then (10) collapses to the conventional intercept shift dummy formulation for modelling the effect of controls. If $k_j=1$ and $d_j=0$ for all j then controls are completely effective in the sense that during each controls period the wage equation becomes a horizontal line at $W_t=W_{jt}^*$.

TABLE 2—NON-LINEAR ESTIMATION OF EQUATION (11)

Parameter (1)	Point Estimate (2)	t-Statistic (3)	Point Estimate (4)	t-Statistic (5)
β_0	3.8	5.1	3.7	5.3
β_1	0.85	11.5	0.84	11.7
β_2	−5.81	−3.5	−4.9	−3.5
β_3	4.5	2.9	4.4	3.0
d_0	−1.2	−2.9	−1.4	−5.0
k_0	0.20	0.7		
d_N	−0.4	−0.3		
k_N	0.58	4.3	0.55	5.7
e_1	−4.5	4.5	4.7	5.3
e_4	0.8	1.2	0.8	1.3
R^2	.90		.90	
d	2.10		2.07	

Using the controls-off wage equation given in equation (1), denoting the guideposts by $j=0$, the four phases of the Nixon controls by $j=N$, and including explosions dummies for Phase I (E_{1t}) and Phase IV (E_{4t}), equation (10) becomes[7]

$$(11)\quad W_t=\beta_0+\beta_1 PE_t+\beta_2 VIN_t+\beta_3 VIN_{t-1}$$

$$+d_0D_{0t}+k_0D_{0t}W_{0t}^*-k_0\beta_0D_0$$

$$-k_0\beta_1D_0PE_t-k_0\beta_2D_0VIN_t-k_0\beta_3D_0VIN_{t-1}$$

$$+d_ND_{Nt}+k_ND_{Nt}W_{Nt}^*-k_N\beta_0D_N$$

$$-k_N\beta_1D_NPE_t-k_N\beta_2D_NVIN_t$$

$$-k_N\beta_3D_NVIN_{t-1}+e_1E_{1t}+e_4E_{4t}+\varepsilon_{11t}$$

The results of estimating (11) over the full-sample period subject to the implied non-linear constraints are given in the second and third columns of Table 2. The results are consistent with the evidence from equations (6) and (7) above. For the guideposts, d_0 is significant and k_0 is not, indicating that the effect of controls during that period may be described by a shift dummy.

[7] On the basis of the previous evidence that there is no significant difference between the rotation equations for the phases of Nixon controls, the Nixon controls are denoted by k_N and d_N where $k_j=k_N$ and $d_j=d_N$ ($j=1,2,3,4$).

For the Nixon controls, k_N is significant and d_N is not, indicating support for the proportional rotation hypothesis.

Setting the insignificant parameters k_0 and d_N equal to zero and reestimating gives the preferred equation for explaining the rate of wage change over the full sample period. The results are presented in the fourth and fifth columns of Table 2. Using this formulation, 90 percent of the variation in the rate of wage change in the full sample period can be explained and the Durbin-Watson statistic d gives no indication of autocorrelation. The estimates of β_0, β_1, β_2, and β_3 are statistically significant and similar in magnitude to the estimates from the controls-off equation.

For the guideposts, the coefficient of the intercept dummy d_0 is significant and indicates a reduction in the rate of wage change of 1.4 percentage points. For the Nixon controls, k_N is significant and indicates that during the four phases of the Nixon controls the rate of wage change was restrained by an amount equal to approximately half of the difference between the guideline and the rate of wage change that would have occurred in the absence of controls.

The evidence on the explosions is similar to that previously obtained—the results indicate a statistically significant explosion following Phase I of 4.7 percentage points and an explosion following Phase IV which is of a smaller magnitude and significant at the .10 but not the .05 level.

The above estimates using the rotation hypothesis for the Nixon controls have three main advantages over the results using shift dummies given in equation (3). First, the use of the rotation hypothesis for the Nixon controls results in a slight increase in explanatory power of the equation with fewer parameters estimated. Second, the estimates using the rotation hypothesis do not require the unwarranted assumption that controls did not affect the slope coefficients of the wage equation during the Nixon controls. Third, and most important, the rotation hypothesis permits an explanation of why the various phases of the Nixon controls differed substantially in the amount of restraint on the rate of wage change.

The amount of restraint during the various periods of controls implied by the preferred estimate of equation (11) is shown in Table 3. The effect of controls (col. (8)) is calculated as the difference between the predicted rate of wage change with $d_0 = k_N = e_1 = e_4 = 0$ (col. (2)) and the predicted rate of wage change with these parameters at their estimated values (col. (6)). The close correspondence between the predicted rate of wage change during controls (col. (6)) and the actual rate of change during controls (col. (7)) illustrates the power of the model to predict during the controls-on period.

The estimated effect on the rate of wage change during the various phases of controls is very similar to the effect indicated by the analysis of residuals given in Table 1. As indicated above, however, the use of the rotation hypothesis permits an explanation of why the various phases of the Nixon controls differed in the amount of restraint on wages. In all four phases the reduction was approximately 0.55 of the amount of the excess over the guideline. During the Phase I freeze, the "zero" guideline was substantially below the predicted rate of increase of 6.6 percent. Elimination of 0.55 of this excess reduced the rate of wage change by 3.6 percentage points, a substantial reduction.

For Phase II, on the other hand, the rotation model indicates that the reason for the lack of restraint was that the mean predicted rate of increase in the absence of controls was only 0.1 percentage points above the 5.5 percent guideline. Thus, even though controls eliminated 0.55 of this excess, there was virtually no restraint on the rate of wage change.

Phases III and IV are both intermediate cases between Phases I and II. For Phase III the predicted rate of wage change in the absence of controls exceeded the 5.5 percent guideline by 0.7 percentage points. Eliminating 0.55 of this excess resulted in a small reduction in the rate of wage change of 0.4 percentage points.

During Phase IV the relatively tight labor market conditions and a rise in the expected inflation rate increased the predicted rate of

TABLE 3—EFFECT OF CONTROLS ON THE RATE OF WAGE CHANGE

Controls Period (1)	Predicted Wage Change Without Controls (2)	Guideline (3)	Excess Over Guideline (4)	Effectiveness Coefficient (5)	Predicted Wage Change During Controls (6)	Actual Wage Change During Controls (7)	Effect of Controls on the Rate of Wage Change (8)
Guideposts	4.2	3.2	1.0	–[a]	2.8	2.8	−1.4
Phase I	6.6	0.0	6.6	0.55	3.0	3.0	−3.6
Phase I Explosion	4.8	–[a]	–[a]	–[a]	9.9	9.9	+5.1
Phase II	5.6	5.5	0.1	0.55	5.6	5.7	0.0
Phase III	6.2	5.5	0.7	0.55	5.8	5.9	−0.4
Phase IV	9.2	5.5	3.7	0.55	7.2	7.2	−2.0
Phase IV Explosion	10.6	–[a]	–[a]	–[a]	11.5	11.5	+0.9

Notes: Dates for the various periods are given in Table 1. The quarter 1972:I is included in the Phase I explosion rather than in the Phase II controls period. The effect of controls given in col. (8) is calculated as the difference between the predicted wage change without controls given in col. (2) and the predicted effect with controls given in col. (6)

[a] Indicates nonapplicable.

wage change in the absence of controls to 9.2 percent; 3.7 percentage points in excess of the 5.5 percent guideline. Elimination of .55 of this excess resulted in a 2.0 percentage point reduction in the rate of wage change. The resulting 7.2 percent mean rate of wage change during Phase IV was, however, still substantially above the guideline.

The results also indicate that the two phases which involved substantial reductions in the rate of wage change, Phase I and Phase IV, were followed by explosions. As before, the magnitude of these explosions was such that it was not possible to reject the hypothesis that the reductions during the Nixon controls were completely offset by the subsequent explosions.[8]

The reason for the explosion following Phase I is obvious—the freeze merely delayed settlements until the end of the ninety-day period. It is less clear why a wage explosion occurred following Phase IV but not following the guideposts. One hypothesis is that an explosion will occur if controls restrain wages more than prices, causing the real wage to be depressed below its equilibrium level, but no explosion will occur if wages and prices are restrained equally. A cursory examination indicated that real wages appeared to be reduced during Phase IV but not during the guideposts, which is consistent with the hypothesis. This is merely a preliminary assessment, however, and the topic requires further research.

IV. Conclusions and Policy Implications

The empirical evidence presented in this paper on the effect of controls on the rate of wage change indicates considerable diversity among the five controls periods examined—the guideposts period and the four phases of the Nixon controls. The guideposts policy appears to have shifted the wage equation downward by 1.4 percentage points, resulting in a reduction in the rate of

[8] Similar conclusions are obtained using the Livingston expectations data. The results indicate that the guideposts can be described by an intercept shift dummy and the Nixon controls by the proportional rotation hypothesis, supporting the results obtained using the rational expectations variable. Using the estimated coefficients to calculate the effect of controls on the rate of wage change, the mean residuals for the guideposts and the four phases of the Nixon controls are −1.1, −3.0, −0.4, −0.5, and −1.0. The results also indicate significant explosions following Phase I and the termination of Phase IV. Again, the hypothesis that these explosions completely offset the reduction while the Nixon controls are in force cannot be rejected.

wage change of that amount. There was no indication of an explosion in wages following the termination of the guideposts policy.

The evidence indicates that the Nixon controls rotated the wage equation at the guideline, that is, they eliminated 0.55 of the difference between the guideline established for wage increases and the rate of change which would have occurred in the absence of controls. The implied amount of the reduction in the rate of wage change, however, differed among the four phases: 3.6 percentage points for Phase I, 0.0 percentage points for Phase II, 0.4 percentage points for Phase III, and 2.0 percentage points for Phase IV. In addition, the two phases which showed substantial reductions, Phase I and Phase IV, were followed by explosions which eliminated most of the effect while controls were in force. Thus, on the basis of the evidence examined, only one of the five periods of controls was successful in the sense of achieving a sustained reduction in the rate of wage inflation. The other four controls periods were unsuccessful even in this limited sense.

In addition to helping explain the varying amounts of restraint in the four phases of the Nixon controls, the rotation hypothesis also helps to understand the rather different assessment of the effects by the Council of Economic Advisors (CEA) who concluded that the evidence "...does support a partial but important judgement about the experience with the controls system: regardless of the overall effect of the program, whatever contribution it may have made was probably concentrated in its first 16 months, when the economy was operating well below its potential" (1975, p. 228). The CEA's judgement regarding the success of the program appears to be based on the fact that the rate of wage change was near the established guideline during Phase II whereas it exceeded the guideline in Phase IV. However, the analysis in this paper indicates that during Phase II the target would have been virtually achieved even without controls, that is, there was no independent effect of controls. During Phase IV, on the other hand, although the rate of change was in excess of the 5.5 percent guidelines, it was substantially below the rates which have occurred without controls.

Controls could be a useful tool of macroeconomic policy if they were used to reduce expected inflation indirectly by forcing down actual rates of wage and price increase. Controls could achieve such a reduction with a less severe rise in unemployment during the "short-run" transition period than would occur if only monetary and fiscal policy were used to restrain the economy.

If controls are to be successful, however, they must be accompanied by a monetary and fiscal policy which is consistent with the lower rates of wage and price change forced on the economy with controls. If policymakers attempt to maintain a disequilibrium position with controls, rather than speeding the transition to a new equilibrium, the result will very likely be an explosion in the rates of wage and price change when controls are removed. There is some indication that a successful controls program must also affect wages and prices equally so that the real wage is not displaced from its equilibrium level during controls. Finally, a successful controls policy requires that the wage guideline be chosen with some care to ensure that it is below the rate of wage change which would have occurred without controls.

Appendix: Data Sources

The source of the wage index is: January 1959–December 1974, *The Hourly Earning Index, 1964–August 1975*; September 1975–March 1978, *Monthly Labor Review*, various issues. The Livingston data are the revised twelve-month forecasts given in Carlson, semiannual, interpolated to quarterly. The *CPI* is All Urban Consumers, all items, from the *Monthly Labor Review*, various issues. Data for the civilian labor force are the revised figures, January 1959–December 1966, *Employment and Earnings*, February 1973 issue; January 1967 to December 1974, February 1975 issue; January 1975 to March 1978, various issues.

One problem in using a quarterly model is that the beginning and ending months of the controls periods do not correspond exactly to quarters. This problem was not regarded as serious, however, except for Phase I, the ninety-day freeze which extended from August 15, 1971 to November 15, 1971, that is, from the middle of the third quarter to the middle of the fourth quarter of 1971. To obtain a more accurate measurement of the effects of Phase I and any subsequent explosion, the observations on the dependent and independent variables for 1971:IV and the following two quarters were replaced by values calculated from monthly data corresponding to the three months of the Phase I freeze and the following two three-month periods. Reverting to the normal 1972:I observation would have resulted in the omission of the month of December 1971; a month which it was felt on a priori grounds might contain a substantial amount of wage explosion due to the wage increases deferred during Phase I.

REFERENCES

S. W. Black and H. H. Kelejian, "A Macro Model of the U.S. Labor Market," *Econometrica*, Sept. 1970, *38*, 712–41.

Ronald G. Bodkin et al., *Price Stability and High Employment: The Options for Canadian Economic Policy*, Special Study no. 5, Economic Council of Canada, Ottawa 1967.

J. Carlson, "A Study of Price Forecasts," *Annals Econ. Soc. Measure.*, Mar. 1977, *6*, 27–56.

______ **and M. Parkin,** "Inflation Expectations," *Economica*, May 1975, *42*, 123–38.

G. De Menil and S. S. Bhalla, "Direct Measurement of Popular Price Expectations," *Amer. Econ. Rev.*, Mar. 1975, *65*, 169–80.

O. Eckstein and R. Brinner, "The Inflation Process in the U.S." study for the Joint Economic Comm. 92d Cong. 2d sess. 1972.

E. L. Feige and D. K. Pearce, "Inflation and Incomes Policy: An Application of Time Series Models," in Karl Brunner and Allan H. Meltzer, eds., *The Economics of Price and Wage Controls*, New York 1976.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

R. J. Gordon, "Inflation in Recession and Recovery," *Brookings Papers*, Washington 1971, *1*, 105–58.

______, "Wage-Price Controls and the Shifting Phillips Curve," *Brookings Papers*, Washington 1972, *2*, 387–421.

______, "The Response of Wages and Prices to the First Two Years of Controls," *Brookings Papers*, Washington 1973, *3*, 765–78.

R. G. Lipsey and J. M. Parkin, "Incomes Policy: A Re-appraisal," *Economica*, May 1970, *37*, 115–38.

Timothy W. McGuire, "On Estimating the Effects of Controls," in Karl Brunner and Allan H. Meltzer, eds., *The Economics of Price and Wage Controls*, New York 1976.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

Walter Oi, "On Measuring the Impact of Wage-Price Controls: A Critical Appraisal," in Karl Brunner and Allan H. Meltzer, eds., *The Economics of Price and Wage Controls*, New York 1976.

M. Parkin, M. Sumner, and R. A. Jones, "A Survey of the Econometric Evidence of the effects of Incomes Policy on the Rate of Inflation," in Michael Parkin and Michael Sumner, eds., *Incomes Policy and Inflation*, Manchester 1972.

G. Perry, "Changing Labor Markets and Inflation," *Brookings Papers*, Washington 1970, *3*, 411–41.

Edmund S. Phelps, "Money Wage Dynamics and Labor Market Equilibrium," in his *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, Nov. 1958, *65*, 283–300.

F. Reid, "The Expectations Hypothesis of the Phillips Curve and the Rotation Hypothesis of Incomes Policy: Empirical Tests and Policy Implications," unpublished doctoral dissertation, Queen's Univ. 1975.

N. J. Simler and A. Tella, "Labor Reserves and the Phillips Curve," *Rev. Econ. Statist.*, Feb. 1968, *50*, 32–49.

S. Vroman and W. Vroman, "Money Wage

Changes: Before, During and After Controls," *Southern Econ. J.*, Apr. 1979, *45*, 1172–87.

The Conference Board, *The Conference Board Statistical Bulletin*, New York, various issues.

U.S. Bureau of Labor Statistics, *Employment and Earnings*, various issues.

______, *The Hourly Earning Index, 1964–August 1975*, Bull. 1897, Washington 1976.

______, *Mon. Labor Rev.*, various issues.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington, various years.

The Duration of Unemployment and Unexpected Inflation: An Empirical Analysis

By Anders Björklund and Bertil Holmlund*

A distinguishing message of the theory of search unemployment is that short-run unemployment fluctuations are explainable by inflationary surprises. Unemployment is basically viewed as productive investment in job search, chosen by employees in order to enhance their lifetime earnings. An increase in aggregate demand will imply a temporary fall in unemployment due to short-run deviations between actual and expected wages; workers are fooled into accepting more employment.

This information-lag interpretation of changes in unemployment might be compared to an alternative view, where the quantity-rationing rules of the labor market are emphasized. A rising flow of labor from unemployment is, according to this theory, caused by the relaxation of job-rationing constraints rather than unanticipated inflation.

In this paper we address ourselves to the question of the empirical importance of the two competitive explanations. The two stories are, of course, not mutually exclusive; we try, via a fairly simple specification, to capture both views in one equation. The principal contribution of our study lies in its ability to provide information about the relative importance of unexpected inflation and job opportunities as explanations of the duration of unemployment.[1] Another interesting feature of our paper is its comparative perspective; we apply the same model to both Swedish and *U.S.* data, thereby being able to reveal certain important differences between the labor markets in the two countries. We find, for example, perhaps somewhat surprisingly, that the *U.S.* unemployment duration is more or less unaffected by unexpected inflation, whereas the results for Sweden, on the other hand, give some support for the information-lag hypothesis. A third novelty of our study is the disaggregated data used (for Sweden only). By focusing the analysis on transition probabilities for workers with different lengths of (incompleted) spells, some interesting behavioral differences are observed; one finding is that the simple information-lag story is more valid for the short-term unemployed.

The paper is organized as follows: Section I introduces the basic theoretical framework that guides our empirical estimation procedures; the latter are described in Section II. Section III presents the data employed and Section IV the empirical results. Some interpretations of our findings are discussed in the final section.

I. Optimal Search Policies and the Duration of Unemployment

Micro-economic explanations of unemployment have focused on the behavior of the household, whereas the demand side generally has been considered as exogenous. We will follow that partial-equilibrium approach, using a simple job search model as our theoretical framework.

Consider the behavior of an unemployed worker according to the search model. His problem is to choose an acceptance wage which assures him an income greater than what he might have received by continued search. The decision is affected by the perceived location of the wage offer distribution. If a monetary contraction produces a leftward shift of the wage offer distribution —or a lower rate of wage inflation—this

*Stockholm School of Economics and The Industrial Institute for Economic and Social Research, Stockholm, respectively. We are indebted to George Borts, Ned Gramlich, Mats Persson, and an anonymous referee for helpful comments on an earlier version.

[1] The question has earlier been addressed by John Barron and Putte Axelsson and Kalle Löfgren. Their methods differ from ours.

change in general market conditions is assumed to be imperfectly detected by job seekers, who mistakenly blame local circumstances rather than changes in aggregate demand. Unemployed workers will search for a longer time causing the length of spells of unemployment to rise.

A common assumption in standard search models is that the number of job offers received per period equals one. The probability of leaving unemployment—the transition probability—is then solely determined by the job seeker's offer-acceptance probability. The simplifying job offer assumption is, however, not inherent in search theory per se; by a modest generalization the case with random number of job offers is easily incorporated into the basic search-theoretic framework. Consider the job seeker's transition probability, which—in the absence of labor force exits—equals the hiring probability. Decomposing the transition probability (μ) into two components, the job offer probability (θ) and the acceptance probability (P) we have

$$(1) \qquad \mu=\theta P=\theta[1-F(a)] \qquad \theta\leqslant 1$$

where a is the reservation wage and $F(\cdot)$ the distribution function of wage offers. If the transition probability is constant during search, the expected duration of unemployment (D) is

$$(2) \qquad D=1/\mu=1/\theta[1-F(a)]$$

What then are the characteristics of an optimal search policy? In the simple case of infinite time horizon and discount rate r, the optimal policy implies a certain time invariant reservation wage obtained as the solution to

$$(3) \qquad C+a=\frac{\theta P}{r}[E(w|w>a)-a]$$
$$=\frac{\theta}{r}\int_a^\infty (w-a)f(w)\,dw$$

where C is the (constant) marginal search cost and $f(\cdot)$ the known density function of wage offers.[2] Equation (3) implies that the reservation wage declines as the job offer probability θ decreases. Likewise, a known leftward shift of the wage offer distribution will also reduce the reservation wage.

We have so far briefly outlined the basic search story, strictly valid only in a stationary world. Now consider the possibility of fluctuations in aggregate demand, influencing the job seeker's transition probability via the job offer probability (more vacancies) and/or via imperfect reservation wage adjustments. Three different effects may be identified:

1) *The pure availability effect.* An increasing number of vacancies means a higher job offer probability, thereby reducing the duration of unemployment.

2) *The supply effect.* A permanent increase of the job offer probability will increase the expected returns from search, thus increasing the worker's reservation wage. It follows that the unemployment effect of a rising number of vacancies is ambiguous a priori. Robert Feinberg has, however, demonstrated that the availability effect will outweigh the supply effect under certain reasonable assumptions.

3) *The detection-lag effect.* Changes in aggregate demand will affect the location of the wage offer distribution. Assuming a lag in the discernment of a rising rate of inflation, reservation wages will be unaffected in the short run, implying a rising flow of new hires from the pool of unemployed.

Summarizing these three effects we have

$$(4) \qquad \mu=\theta\Big(\underset{+}{V}\Big)P\Big(\underset{-}{V},\underset{+}{w/w^*}\Big)=g(V,w/w^*)$$

where V is the number of vacancies, w the actual average wage, and w^* the expected average wage.

We would argue that equation (4) represents the kernel of the search theory of cyclical unemployment. The standard search model outlined does rely on some very restrictive assumptions, for example, a stationary wage offer distribution, fixed leisure time, and a constant job offer probability. More complex search models, for example, those of Claes-Henric Siven and John Seater (1977, 1978, 1979) are, however, fairly con-

[2] For a proof of (3), see, for example, Steven Lippman and John McCall.

sistent with the simple search model in their emphasis on unexpected inflation and vacancy contacts.[3] We are suppressing other plausible determinants of unemployment duration, for example, variations in unemployment compensation and the discount rate. These simplifications should not be too severe, since the cyclical fluctuations are dominating in the data. We have also excluded changes in the price level from consideration, perhaps a more questionable simplification. Unexpected *price* inflation does affect unemployment in some models within the micro foundations literature, although it is absent in the standard search model. The interpretation of this candidate regressor is, however, quite different in, for example, the Lucas-Rapping model compared to the Siven model (misperception of *future* prices vs. misperception of *current* prices) and the theoretical predictions are completely opposite; a higher rate of unexpected price inflation will *increase* unemployment in Siven's model and *decrease* unemployment in the Lucas-Rapping model.[4] It is also interesting to note an important result from Seater's "unified model": unexpected wage changes affect unemployment duration irrespective of how workers perceive accompanying price changes.[5] We decided to exclude the price inflation variable from the regressions, thereby avoiding troublesome problems of interpretation.

II. Empirical Analysis

A straightforward method of investigating the validity of the detection-lag hypothesis is to specify explicit transition probability equations with vacancies and unexpected wage increases as explanatory variables, that is, to represent equation (4) above by a suitable functional form. The basic specification used will be

$$(5) \quad ln\mu_t = \alpha_1 + \alpha_2 lnV_t + \alpha_3 ln(w_t/w_t^*)$$

The obtained α_2 estimate reflects the net result of the positive availability effect and the negative supply effect; intuition and some theoretical predictions suggest that α_2 (the net availability effect) will have a positive sign.[6]

The main problem with the approach chosen is, of course, that it requires an analysis of perceived as well as actual wages. Since no direct data about expected wages or wage changes are available, some model of the formation of expectations must be used. The expanding literature about the formation of expectations give several alternatives which all are quite plausible. However, no model which can be made operational can be considered "correct" in all

[3] The worker in Siven's and Seater's models is maximizing his lifetime utility by using search in the labor market as one important choice variable. Siven also considers search in the goods market but assumes leisure to be fixed; maximization of the utility function is therefore equivalent to maximization of lifetime earnings. Seater, on the other hand, takes account of variable leisure but ignores search in the goods market.

[4] Unexpected price inflation in the Siven model implies a reallocation of time from search in the labor market to search in the goods market thereby causing a decline of the job offer probability. The reservation wage will also increase, reinforcing the effect on unemployment duration. The Lucas-Rapping model is hardly suitable for analyzing the length of spells of unemployment since it disregards job search and considers unemployment as pure leisure, resulting as a difference between actual and normal employment. Michael Darby and Jonathan Kesselman and N. Eugene Savin have run unemployment regressions for the United States including unanticipated price increases as an explanatory variable. The results turn out to be unsatisfactory; the coefficients are as a rule insignificantly different from zero and the signs are unstable across different regressions.

[5] See Seater (1978).

[6] The crucial element in Barron's approach—followed by Axelsson and Löfgren—is to construct a model which gives an explicit specification of the relationship between the number of vacancies (V) and the job offer probability (θ). Given such a relationship, $\theta = f(V)$, the acceptance probability is obtained as $P = \mu/f(V)$. The procedure is interesting since it can validate a procyclical reservation wage pattern (i.e., P and V are inversely correlated). The approach requires, however, some fairly restrictive assumptions regarding the relationship between θ and V; Barron assumes that $\theta = k \cdot V$, implying that the elasticity $\partial ln\theta/\partial lnV$ equals one, an implication from the assumption that each firm has only one vacancy in each occupation. It can be shown that less restrictive assumptions produce an elasticity lower than one. Barron's procedure is, moreover, unable to separate the supply effect from the detection-lag effect. Our approach, on the other hand, can quantify the detection-lag effect but captures only the net availability effect.

respects. Our approach has been to try three different models in order to investigate how robust the information-lag hypothesis is with respect to the different specification. Two of the applied forecasting functions are consistent with the idea that workers learn from past errors, reestimating the parameters of their forecasting equations when more information is obtained.

A. *Adaptive Expectations*

The first model used is a type of adaptive expectations. These expectations are formed according to a finite distributed lag of past wage changes, that is, with quarterly data (which is used for Sweden):

$$\text{(6a)}\qquad \left(\frac{w_t^*}{w_{t-4}}\right)=\sum_{i=1}^{4} l_i\left(\frac{w_{t-i}}{w_{t-4-i}}\right)$$

where

$$\text{(6b)}\qquad \sum_{i=1}^{4} l_i=\frac{1}{10}\sum_{i=1}^{4}(5-i)=1$$

and with monthly data (which is used for the United States):

$$\text{(7a)}\qquad \left(\frac{w_t^*}{w_{t-12}}\right)=\sum_{i=1}^{12} l_i\left(\frac{w_{t-i}}{w_{t-12-i}}\right)$$

where

$$\text{(7b)}\qquad \sum_{i=1}^{12} l_i=\frac{1}{78}\sum_{i=1}^{12}(13-i)=1$$

Models like these—where the sum of the weights has been constrained to one—are often used in empirical work even though it has been pointed out that the theoretical basis is quite weak. (See, for example, Mats Persson, where it is shown that the sum should equal one only in very special cases if the forecast is to be optimal.)

B. *Expectations from an* ARMA *Process*

Even though the simplicity of the simple adaptive model is appealing—since it might be argued that workers form their expectations in a simple and cheap way—it could also be argued that individuals have some knowledge about historical regularities of wage changes, and that they use this information when forming their expectations. One possible way to represent these regularities is to apply a time-series approach. The assumption is that people have in their mind an autoregressive moving-average process (ARMA) which is generating forecasts from period to period. Both the specification and the parameters of this process are, however, likely to be revised when people receive more information about wage changes. Therefore we have proceeded as follows: The process has been reestimated each period and reidentified each fourth period (with quarterly data) and each twelfth period (with monthly data).[7] For Sweden, the character of the process changed over time; when observations from 1960 onwards were used the appropriate process changed from an $AR(1)$ to an $AR(1)MA(2)$, back again to an $AR(1)$ and finally—during the past two years (1976–77)—an $MA(10)$ on the first differences of the variable (i.e., the process was nonstationary). All the time autoregressive seasonal terms had to be used.

For the United States, the process was stationary when data from 1960 to 1969 were used—$AR(1)$ with first a seasonal autoregressive term and then a seasonal moving-average term. From then on the process became nonstationary with an $MA(1)$ term and a seasonal moving-average term on the first differences.

C. *Expectations from an Estimated Wage Equation*

It could, finally, be argued that workers are still more rational than using information only from an ARMA process of wage changes. They might even have in mind an empirical model incorporating different economic variables. An unemployed worker forming his expectations may, for example, use a wage equation of the Phillips curve

[7] A Box-Jenkins program called *T*-series available at the Stockholm School of Economics has been used. For identification criteria, see Charles Nelson.

type. We have therefore estimated wage equations (quarterly data) as:

$$(8)\quad WCH_t = \beta_1 + \beta_2 \cdot (V_{t-1} + V_{t-2} + V_{t-3} + V_{t-4}) + \beta_3 WCH$$

where $WCH = (w_t - w_{t-4})/w_{t-4}$

Again, the model was reestimated each fourth period and with data from the last five years. On the whole, the estimated equations performed reasonably well for Sweden according to standard statistical criteria. This approach was less successful for the United States; the available vacancy indicators turned out to be bad predictors of wage inflation. We decided to exclude this expectations-formation scheme for the *U.S* regressions.

III. The Data

Swedish transition probabilities have been estimated as follows: The rotating system of the "Swedish Labor Force Surveys" is constructed so that almost 90 percent of those who are interviewed in one survey are interviewed again three months later, whereas different individuals are interviewed in two subsequent months. In order to improve the estimates we decided to compute quarterly transition probabilities.

Denoting the number of unemployed for at least *a* weeks but less than *b* weeks at time t by $G_t^{a,b}$, and the weekly inflow into unemployment by f, we can describe the estimates as follows:

$$(9)\quad G_t^{1,14} = f \sum_{i=0}^{12} (1-\mu_1)^i$$

$$(10)\quad G_{t+13}^{14,27} = G_t^{1,14}[1-\mu_2]^{13}$$

$$(11)\quad G_{t+26}^{27,39} = G_{t+13}^{14,27}[1-\mu_3]^{13}$$

Three transition probabilities are obtained (μ_1, μ_2, and μ_3) which can be regarded as conditional upon the length of the spell of unemployment. By using available data on f, $G_t^{1,14}$, $G_{t+13}^{14,27}$, etc., we obtain μ_1 from

$$(12)\quad \frac{(1-\mu_1)^{13}-1}{(1-\mu_1)-1} = \frac{G_t^{1,14}}{f}$$

whereas μ_2 and μ_3 are calculated as

$$(13)\quad \mu_2 = 1 - \left[\frac{G_t^{14,27}}{G_t^{1,14}}\right]^{1/13}$$

$$(14)\quad \mu_3 = 1 - \left[\frac{G_{t+26}^{27,39}}{G_{t+13}^{14,27}}\right]^{1/13}$$

The Swedish vacancy statistics are from "Labor Market Statistics," published by Arbetsmarknadsstyrelsen (the National Labor Market Board). Quarterly wage data are obtained from the labor market issues of Statistical Reports, published by Statistiska Centralbyrån (the National Bureau of Statistics). All data used refer to manufacturing industry.

The *U.S.* transition probabilities refer to the labor market as a whole. They were computed by using the method proposed by Barron. The essential idea is to compare the number of people in one week who have been unemployed less than five weeks with the number of people four weeks later who have been unemployed five to eight weeks. The difference consists of people who have left the pool of unemployed. The duration data reported in *Employment and Earnings* are grouped in the classes one to four weeks, five to fourteen weeks, etc., which requires a slight modification of the method outlined above; for details, see Barron.

The *U.S.* wage data are average hourly earnings in manufacturing industry, reported in *Employment and Earnings*.[8] As vacancy data for the period 1965–75 we used the Help-Wanted Advertising Index (*HWA*) published in *Main Economic Indicators*. For the period 1969.4–1973.10, manufacturing vacancies (*Vm*) according to establishment data were also tried (see

[8] In some regressions we also tried average hourly earnings for the total private nonagricultural sector. The results were basically the same.

Employment and Earnings); the latter series are available only for (approximately) this period.

IV. Empirical Results

The results from alternative estimations are presented in Tables 1 and 2. The estimation method is weighted least squares and the appropriate weights are derived in an appendix which is available from us upon request.

Let us first look at the results obtained for Sweden, shown in Table 1. We observe, in the first place, that the detection-lag variable is significant both for the short-term unemployed (one to thirteen weeks) and for the medium-term unemployed (fourteen to twenty-six weeks). These results hold for all models of expectations.[9] For the long-term unemployed, on the other hand, no significant detection-lag effect is revealed; the coefficient has even a wrong sign. The job availability variable (V) is significantly positive in all regressions, even for the long-term unemployed. Dropping this variable produces in most cases a marked decrease in the *D.W.* value, indicating the presence of specification errors.

What then are the economic interpretations of the different results for the three groups of unemployed? No straightforward answer is available, partly because the "hypothesis-testing includes a joint test of the underlying model and the expectations-generating mechanism" (see Anthony Santomero and Seater, p. 525). The absence of any significant detection-lag effect for the long-term unemployed may have at least two explanations. There are arguments in favor of both these interpretations. First, it makes sense to hypothesize that the long-term unemployed (more than six months in our data) are better informed about the actual wage offer distribution, simply because they have experienced a longer period of "learning" through full-time job search. This argument implies that the parameters of the forecasting function might differ across workers with different unemployment histories.

The second interpretation may be elucidated by recalling some familiar results from search theory: The reservation wage of a job seeker with finite search horizon will, under some stationary conditions, fall with the duration of unemployment, a theoretical prediction which has been given empirical support.[10] Eventually the reservation wage will coincide with the minimum value of the wage offer distribution, implying an acceptance probability equal to one. In that extreme case, all job offers are accepted and there is no detection-lag effect.

Both of the hypotheses outlined are consistent with the results obtained. Intuition would suggest that both of the mechanisms are in operation to some extent, reinforcing each other and thereby producing the observed results.

Since both the (net) availability effect and the detection-lag effect are significant, it is important to find out the relative importance of these variables as determinants of the cyclical variations of the duration of unemployment. To do this we must take the *size* of the parameters as well as the *variation* of the independent variables into account. The question might be illuminated by comparing the predicted transition probabilities using estimates from regressions in the table:

$$\hat{\mu}_t = \alpha_1 \cdot V_t^{\alpha_2} \cdot \left(\frac{w_t}{w_t^*}\right)\alpha_3 \tag{15}$$

with the transition probabilities obtained when inflation is perfectly foreseen ($w_t = w_t^*$):

$$\tilde{\mu}_t = \alpha_1 \cdot V_t^{\alpha_2} \tag{16}$$

Using the results from the adaptive model Figure 2 demonstrates the relative unimportance of the detection-lag effect for the

[9] We have also tried logit specifications in some cases, as well as adaptive expectations with shorter lags. The results turned out to be fairly robust with respect to these changes.

[10] See articles by Reuben Gronau, Hirschel Kasper, and Nicholas Kiefer and George Neumann.

TABLE 1—TRANSITION PROBABILITY EQUATIONS FOR SWEDEN
(Quarterly data 1968.1–1977.3)

	V	w/w^*	$\bar{R}^2$	$D.W.$
Adaptive Expectations				
Short-Term Unemployed (μ_1)				
(1)	0.81	10.30	0.60	1.75
	(4.29)	(4.10)		
(2)	1.11	–	0.42	1.57
	(5.36)			
(3)	–	14.51	0.41	1.05
		(5.18)		
Medium-Term Unemployed (μ_2)				
(4)	0.34	1.97	0.33	2.27
	(3.24)	(1.69)		
(5)	0.41	–	0.30	2.29
	(4.11)			
(6)	–	3.42	0.16	1.84
		(2.83)		
Long-Term Unemployed (μ_3)				
(7)	0.39	−3.35	0.09	2.16
	(2.19)	(−1.57)		
(8)	0.31	–	0.05	2.28
	(1.78)			
(9)	–	−1.99	0.004	2.03
		(−0.93)		
ARMA Expectations				
Short-Term Unemployed (μ_1)				
(10)	0.98	7.47	0.50	1.43
	(4.94)	(2.59)		
(11)	–	11.00	0.18	0.79
		(3.08)		
Medium-Term Unemployed (μ_2)				
(12)	0.36	2.21	0.33	2.19
	(3.56)	(1.64)		
(13)	–	3.56	0.11	1.72
		(2.41)		
Long-Term Unemployed (μ_3)				
(14)	0.36	−2.16	0.05	2.27
	(1.94)	(−0.81)		
Expectations from Wage Equations				
Short-Term Unemployed (μ_1)				
(15)	1.13	8.43	0.51	1.66
	(5.93)	(2.79)		
(16)	–	7.76	0.06	0.65
		(1.85)		
Medium-Term Unemployed (μ_2)				
(17)	0.40	3.10	0.37	2.27
	(4.23)	(2.38)		
(18)	–	3.37	0.09	1.63
		(2.14)		
Long-Term Unemployed (μ_3)				
(19)	0.30	−3.20	0.07	2.20
	(1.72)	(−1.31)		

Note: $\bar{R}^2$ is the fraction of the weighted variance of the dependent variable explained by the weighted independent variables, adjusted for degrees of freedom. The $\bar{R}^2$ obtained when regressing μ_1 on $\hat{\mu}_1$ from equation (1) was 0.62.

TABLE 2—TRANSITION PROBABILITY EQUATION FOR THE UNITED STATES (Monthly Date 1969.4–1973.10 and 1965.2–1975.12)

	HWA	*Vm*	*w/w**	*TIME*	$\bar{R}^2$	*D.W.*	ρ
Adaptive Expectations							
1969.4–1973.10							
1	–	0.23	1.62	−0.0008	0.73	1.19	–
		(11.27)	(1.59)	(−1.61)			
2	–	0.21	0.91	−0.0002	[a]	2.02	0.29
		(8.57)	(1.11)	(−0.41)			
3	–	0.24	1.42	–	0.72	1.13	–
		(11.81)	(1.38)				
4	–	0.21	0.87	–	[a]	2.03	0.30
		(8.71)	(1.08)				
5	–	–	0.14	−0.0021	0.07	0.37	–
			(0.08)	(−2.38)			
6	0.50	–	1.36	−0.0031	0.72	1.18	–
	(11.20)		(1.33)	(−6.39)			
7	0.44	–	0.71	−0.0022	[a]	1.97	0.31
	(8.21)		(0.86)	(−3.84)			
8	0.45	–	0.21	–	0.51	0.67	–
	(7.64)		(0.15)				
1965.2–1975.12							
9	0.52	–	0.70	−0.0025	0.83	1.34	–
	(16.81)		(1.28)	(−19.74)			
10	0.53	–	0.47	−0.0025	[a]	2.03	0.34
	(11.45)		(0.81)	(−13.26)			
11	0.49	–	1.55	–	0.33	0.34	–
	(7.93)		(1.43)				
12	–	–	0.71	−0.0024	0.47	0.44	–
			(0.73)	(−10.62)			
ARMA Expectations							
1969.4–1973.10							
13	–	0.23	2.57	−0.0007	0.74	1.15	–
		(11.27)	(2.07)	(−1.54)			
14	–	0.20	1.96	−0.0002	[a]	2.03	0.30
		(8.63)	(1.98)	(−0.40)			
15	–	0.24	2.48	–	0.73	1.10	–
		(11.92)	(1.98)				
16	–	0.20	1.93	–	[a]	2.04	0.31
		(8.81)	(1.98)				
17	–	0.24	–	–	0.71	1.12	–
		(11.64)					
18	–	0.20	–	–	[a]	2.04	0.30
		(8.68)					
19	–	–	3.03	−0.0021	0.10	0.36	–
			(1.31)	(−2.49)			
20	0.49	–	2.48	−0.0030	0.73	1.17	–
	(11.23)		(2.00)	(−6.44)			
21	0.44	–	1.81	−0.0022	[a]	1.99	0.30
	(8.36)		(1.80)	(−3.96)			
22	0.44	–	2.23	–	0.53	0.65	–
	(7.71)		(1.34)				
1965.2–1975.12							
23	0.52	–	−0.37	−0.0026	0.83	1.33	–
	(16.7)		(−0.50)	(−19.19)			
24	0.53	–	−0.07	−0.0025	[a]	2.04	0.35
	(11.37)		(−0.10)	(−13.01)			
25	0.49	–	3.45	–	0.35	0.41	–
	(8.03)		(2.48)				
26	–	–	−0.09	−0.0025	0.46	0.44	–
			(−0.07)	(−10.31)			

Note: ρ is the first-order autocorrelation coefficient obtained by using the Cochrane-Orcutt approach.

[a] Not applicable.

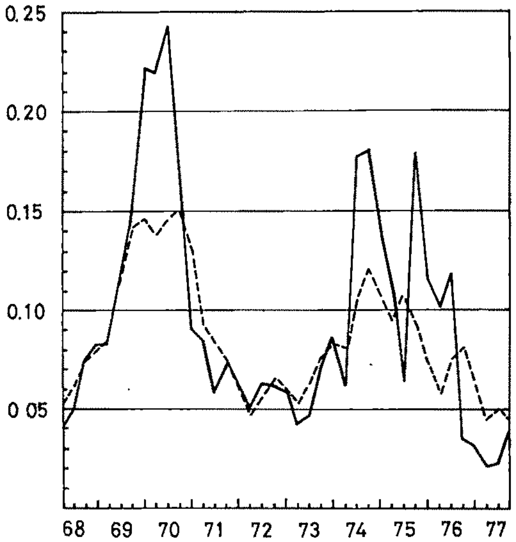


FIGURE 1. THE EFFECTS OF UNEXPECTED INFLATION—SHORT-TERM UNEMPLOYED IN SWEDEN

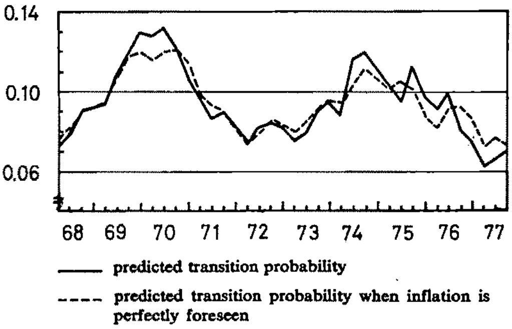


FIGURE 2. THE EFFECTS OF UNEXPECTED INFLATION—MEDIUM-TERM UNEMPLOYED IN SWEDEN

medium-term unemployed. Inflationary surprises produce, on the other hand, quite important unemployment effects for the short-term unemployed during the peak years 1969–70 and 1974–75. (See Figure 1.) The main part of the variation is, however, attributable to the vacancy variable.

Turning now to the *U.S.* regressions shown in Table 2, the dominant availability effect is even more pronounced than in the Swedish case. The vacancy variables used are highly significant in all regressions whereas the detection-lag coefficient is fairly sensitive with respect to the choice of expectations model and estimation period. A significant detection-lag effect is obtained only by applying an ARMA expectations-generating mechanism for the period 1969.4–1973.10. These results are independent of the choice of vacancy variable. Exclusion of the latter also gives rise to a strong decline in the *D.W.* statistic, indicating specification errors. When the estimation period is extended (1965.2–1975.12), the significance of unexpected inflation disappears.[11] It should also be noted that a negative and significant trend coefficient is obtained when *HWA* is used as the vacancy variable.

The main conclusion from these exercises on *U.S.* data is that the job-availability variables are the dominant determinants of the cyclical fluctuations of unemployment duration. We cannot, however, rule out the possibility of some detection-lag effects in operation, at least during certain time periods, especially if the expectations are formed according to an ARMA process rather than adaptively.

V. Concluding Remarks

In job search literature there has been a tendency to overlook the importance of vacancy contacts as determinants of the duration of unemployment, the emphasis instead being placed on inflationary surprises. This (mis)use of the search story does not necessarily follow from the logic of the theory; most search models do recognize the significance of the stream of job offers. The popularity of the detection-lag view is probably its ability to provide a reasonable interpretation of the short-run Phillips curve. The transmission mechanism of aggregate demand policies is explicated in a fairly simple way: an increase in the money growth rate will increase inflation thereby fooling the acceptance decisions of job seekers.

In this paper we have demonstrated that this view has some empirical validity, at least for the short-term unemployed and for a labor market such as Sweden's. But we have also shown that unexpected inflation

[11] The coefficient of w/w^* is significant in equation (25), but the *D.W.* value indicates that the *t*-ratio should not be taken seriously.

can explain only a small part of the actual fluctuations in unemployment duration. Since the flow into unemployment is fairly stable over the cycle, our results imply, moreover, that cyclical changes in the unemployment rate are only slightly affected by inflationary surprises.

The elementary search model—where variations in the job offer probability are disregarded—is then clearly inadequate as an explanation of the short-run Phillips curve. Our results also rule out one of the mechanisms which imply a vertical long-run Phillips curve; the natural rate theory must of course be valid if the detection-lag hypothesis is a sufficient explanation of cyclical changes in unemployment. The results are thus more in accordance with the "mainline" view of inflation and unemployment stressing that aggregate demand influences employment and unemployment via the relaxation of job-rationing constraints rather than via misperceptions of relative wages. It is possible that unanticipated *price* inflation may be of some importance even within the latter framework—as a determinant of the flow of vacancies into the labor market. We are, however, unaware of solid theoretical work on that issue.

Let us, finally, offer some comments on the observed differences between the Swedish and *U.S.* labor markets. Sweden has a highly unionized labor market, and wage bargaining at the national level gives rise to relatively uniform and long-term wage contracts. One would be inclined to expect that this institutional setting would produce fast dissemination of information about the wages in general, thus reducing the importance of information-lag effects. The less-unionized *U.S.* labor market more closely resembles the familiar Phelpsian "island parable" (pp. 6–7), than does the Swedish. The scope for temporary wage misperception would therefore seem to be greater in the United States. In fact, we find the opposite. Why? Let us focus on one additional significant difference between labor market functioning in Sweden and the United States—the importance of temporary layoffs. Temporary layoffs constitute—as Martin Feldstein has pointed out—an important source of *U.S.* unemployment. The *U.S.* manufacturing layoff rate has varied between 10 and 20 percent (of the number of employed workers) per year whereas the corresponding Swedish figures are 2–4 percent. The major part (60–70 percent) of the *U.S.* layoffs are temporary, implying that most workers are ultimately rehired by the same employer. Temporary layoffs in Sweden are, on the other hand, very unusual. Unemployed workers on temporary layoff accounted for 2–3 percent of Swedish unemployment during the period 1975–78. The corresponding *U.S.* figures seem to have fluctuated between 10 and 20 percent.[12] Feldstein's view of those laid off as "waiting" rather than "searching" has been questioned on empirical grounds.[13] The Feldstein hypothesis might, however, be considered as modestly corroborated by our results; one interesting interpretation of our revealed *U.S.*-Sweden differences would be that the extent and intensity of job search among the unemployed is lower in the United States. If unemployed workers on layoff act as if they will be recalled—and therefore abstain from search—there is little scope for detection-lag effects of the traditional type.

A laid-off worker "has a job" in some sense; he is attached to a particular firm and expects to be recalled by his employer. He is probably also well-informed about wage changes in his firm. How then would a nonseeking unemployed worker on layoff respond to unexpected general wage inflation? He would, most likely, be *less* inclined to search, thereby reacting similarly to his employed fellows; a familiar implication of search theory is that quits will *decrease*—via lower propensity to search—as a response to unexpected wage increases. Clearly, temporary layoffs represent a middle state between employment and unemployment. Economic theories designed to explain indi-

[12] For Sweden, see the "Swedish Labor Force Surveys." Feldstein's figures imply that 18 percent of those unemployed in March 1974 were on temporary layoff. The corresponding figure for March 1978 is 11 percent (see *Employment and Earnings*).

[13] See the paper by Thomas Bradshaw and Janet Scholl and the discussion following.

vidual behavior in the polar cases would obviously be less suitable when applied to the middle state.

REFERENCES

R. Axelsson and K. G. Löfgren, "The Demand for Labor and Search Activity in the Swedish Labor Market," *Eur. Econ. Rev.*, Aug. 1977, *9*, 345–59.

J. M. Barron, "Search in the Labor Market and the Duration of Unemployment: Some Empirical Evidence," *Amer. Econ. Rev.*, Dec. 1975, *65*, 934–42.

T. F. Bradshaw and J. L. Scholl, "The Extent of Job Search during Layoff," *Brookings Papers*, Washington 1976, *2*, 515–26.

M. R. Darby, "Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–1941," *J. Polit. Econ.*, Feb. 1976, *84*, 1–16.

M. Feldstein, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers*, Washington 1975, *3*, 725–44.

R. Feinberg, "Search in the Labor Market and the Duration of Unemployment: Note," *Amer. Econ. Rev.*, Dec. 1977, *67*, 1011–13.

R. Gronau, "Information and Frictional Unemployment," *Amer. Econ. Rev.*, June 1971, *61*, 290–301.

H. Kasper, "The Asking Price of Labor and the Duration of Unemployment," *Rev. Econ. Statist.*, May 1967, *49*, 165–72.

J. R. Kesselman and N. E. Savin, "Three-and-a-Half Million Workers Never Were Lost," *Econ. Inquiry*, Apr. 1978, *16*, 205–25.

N. M. Kiefer and G. R. Neumann, "An Empirical Job-Search Model with a Test of the Constant Reservation-Wage Hypothesis," *J. Polit. Econ.*, Feb. 1979, *87*, 89–107.

S. A. Lippman and J. J. McCall, "The Economics of Job Search: A Survey: Part I," *Econ. Inquiry*, June 1976, *14*, 155–89.

R. E. Lucas, Jr. and L. A. Rapping, "Real Wages, Employment and Inflation," in Edmund S. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, London 1971, 257–305.

Charles R. Nelson, *Applied Time Series Analysis for Management Forecasting*, San Francisco 1973.

M. Persson, *Inflationary Expectations and the Natural Rate Hypothesis*, Stockholm Sch. Econ. dissertation 1979.

E. S. Phelps, "Introduction: The New Microeconomics in Employment and Inflation Theory," in his *Microeconomic Foundations of Employment and Inflation Theory*, London 1971, 1–23.

______ **et al.,** *Microeconomic Foundations of Employment and Inflation Theory*, London 1971.

A. M. Santomero and J. J. Seater, "The Inflation-Unemployment Trade-Off: A Critique of the Literature," *J. Econ Lit.*, June 1978, *16*, 499–544.

J. J. Seater, "A Unified Model of Consumption, Labor Supply, and Job Search," *J. Econ. Theory*, Apr. 1977, *14*, 349–72.

______, "Utility Maximization, Aggregate Labor Force Behavior, and the Phillips Curve," *J. Monet. Econ.*, Nov. 1978, *4*, 687–713.

______, "Job Search and Vacancy Contacts," *Amer. Econ. Rev.*, June 1979, *69*, 411–19.

Claes-Henric Siven, *A Study in the Theory of Inflation and Unemployment*, Amsterdam 1979.

Arbetsmarknadsstyrelsen (AMS), "Arbetsmarknadsstatistik," (The National Labor Market Board, "Labor Market Statistics"), Stockholm, various issues.

Organization for Economic Cooperation and Development (OECD), *Main Economic Indicators, 1960–1975*, Paris 1976.

Statistiska Centralbyrån (*SCB*), "Arbetskraftsundersökningar," ("Swedish Labor Force Surveys," *AKU*), yearly averages 1975–78, Stockholm.

U.S. Bureau of Labor Statistics, *Employment and Earnings*, Washington, various issues.

Output, the Stock Market, and Interest Rates

By Olivier J. Blanchard*

This paper develops a simple model of the determination of output, the stock market and the term structure of interest rates. The model is an extension of the *IS-LM* model and borrows from it the assumption that output is determined by aggregate demand and that the price level can only adjust over time to its equilibrium value. However, whereas the *IS-LM* emphasizes the interaction between "the interest rate" and output, this model emphasizes the interaction between asset values and output. Asset values, rather than the interest rate, are the main determinants of aggregate demand and output. Current and anticipated output and income are in turn the main determinants of asset values. It is this interaction that the model intends to capture; its goal is to characterize the joint response of asset values and output to changes in the environment, such as changes or announcement of changes in monetary and fiscal policy. As the above brief description makes clear, anticipations are central to the story; the assumption made in this paper will be one of rational expectations.

The paper is organized as follows. Section I describes the model, and Sections II–IV characterize the behavior of the economy under the extreme but convenient assumption that prices are fixed forever. Sections V and VI extend the analysis to the case where prices adjust over time to their equilibrium value.

I. The Model

Let us assume that the economy is closed and that the physical capital stock is constant. There is one good and four marketable assets. These are shares which are titles to the physical capital, private short- and long-term bonds issued and held by individuals, and outside money.

A. *Equilibrium in the Goods Market*

We shall assume that there are three main determinants of spending. The first is the value of shares in the stock market—the stock market for short; being part of wealth, it affects consumption; determining the value of capital in place relative to its replacement cost, it affects investment[1] (see James Tobin). The second is current income which may affect spending independently of wealth if consumers or firms are liquidity constrained. The third is fiscal policy, both through public spending and taxes; fiscal policy will be summarized by an index rather than by an explicit treatment of both taxes and spending. A change in the index may be thought of as a balanced budget change in public spending. Total spending is expressed as

$$d=aq+\beta y+g; \quad a>0; \quad \beta\varepsilon[0,1)$$

All variables are real, d denotes spending, q is the stock market value, y is income, and g is the index of fiscal policy.

Output adjusts to spending over time:

$$\text{(1)} \qquad \dot{y}=\sigma(d-y) \qquad \sigma>0$$
$$=\sigma(aq+g-by) \qquad b\equiv 1-\beta>0$$

where a dot denotes a time derivative.

There are two interpretations of (1), leading to the same functional form. The first is the one given implicitly above which assumes that inventories are decumulated after

*Harvard University. I thank Rudiger Dornbusch and Larry Summers for useful discussions. Financial assistance from the Alfred P. Sloan Foundation is gratefully acknowledged.

[1] A more detailed specification of aggregate demand would distinguish between the average value of capital which affects consumption and the marginal value which determines investment. It would also possibly introduce human wealth and outside money as other components of wealth.

an increase in aggregate demand until production is increased to meet demand.[2] The second is that spending is always equal to production but that actual spending adjusts slowly to desired spending d. The first emphasizes the costs of adjusting production, the second the slow adjustment of spending. The interpretations are not mutually exclusive.

B. *Equilibrium in the Assets Markets*

The three nonmoney assets are assumed to be perfect substitutes. Hence arbitrage between them implies that they have the same expected short-term rate of return.[3] Their common expected rate of return must in turn be such that agents are satisfied with the proportion of money in their portfolios.

Portfolio balance is characterized by a conventional *LM* relation in inverse form:

$$(2)\qquad i=cy-h(m-p)\qquad c>0;\, h>0]$$

where i denotes the short-term nominal rate, y denotes income, and m and p denote the logarithms of nominal money and the price level. The short-term real rate is defined as

$$(3)\qquad r^*\equiv i-\dot{p}^*$$

An asterisk denotes an expectation; $\dot{p}^*$ is the expected rate of inflation.

1. *Arbitrage between Short- and Long-Term Bonds*

The long-term bonds are consols with yield I and price $1/I$. The expected short-term nominal rate of return from holding consols is therefore

$$I\left(1+\frac{d}{dt}\left(\frac{1}{I}\right)\right)=I-\dot{I}^*/I$$

It is the sum of the yield and the expected nominal capital gain. Arbitrage between short and long bonds implies[4]

$$(4)\qquad I-\dot{I}^*/I=i$$

or equivalently $\dot{I}^*/I=I-i$. If the long-term rate I is above the short-term rate i, agents must be expecting a capital loss on consols, that is, an increase in the long-term rate. Let R be the long-term real rate. Then by an equation similar to (4); we can define it implicitly by

$$(5)\qquad r^*=R-\dot{R}^*/R$$

2. *Arbitrage between Short-Term Bonds and Shares*

As q is the real value of the stock market, the expected real rate of return on holding shares is $\dot{q}^*/q+\pi/q$, where π denotes real profit. Real profit is in turn assumed to be an increasing function of output:

$$\pi=\alpha_0+\alpha_1 y;\quad \alpha_1\geqslant 0$$

Arbitrage between short-term bonds and shares therefore implies

$$(6)\qquad \frac{\dot{q}^*}{q}+\frac{\alpha_0+\alpha_1 y}{q}=r^*$$

Equations (1)–(6) characterize output, the stock market and interest rates as functions of policy variables m and g, expectations $\dot{q}^*$ and $\dot{p}^*$, and the price level p. The system is recursive: long rates are determined by equations (4) and (5) but do not in turn determine other variables. Following Tobin, the only link between assets and goods markets is the value of the stock market, q. To close the model, assume that expectations are formed rationally. This leaves us with the need for only one equation, the equation describing the behavior of the price level.

II. Steady State and Dynamics with Fixed Prices

We start with the assumption that prices are fixed. Hence, there is no actual and no expected inflation; nominal and real rates

[2] A more satisfactory—and more complex— formulation would allow for inventories to be rebuilt later during the adjustment process.

[3] I use "short term" instead of "instantaneous" which is more proper but also more cumbersome.

[4] The implicit assumption is that agents hold their expectations with subjective certainty. If this was not the case, the arbitrage equations would have to pay attention to Jensen's inequality.

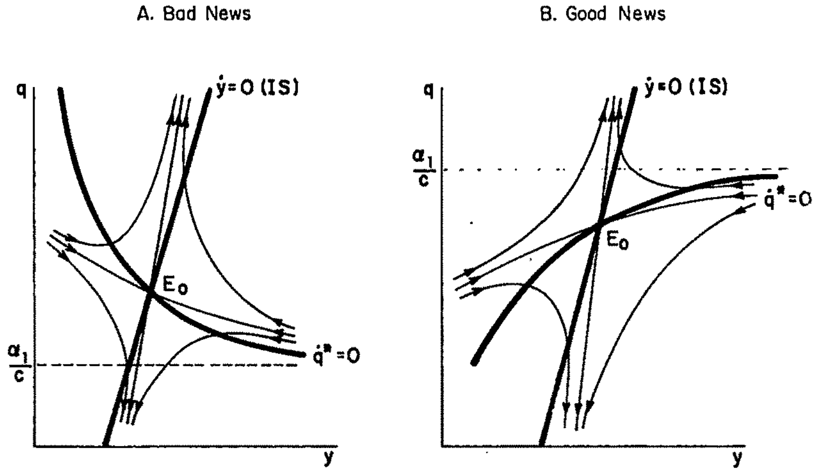


FIGURE 1

are identical and the system simplifies to

(1) $\dot{y}=\sigma(aq-by+g)$

(2′) $r=cy-h(m-p)$

(6′) $\frac{\dot{q}^*}{q}+\frac{\alpha_0+\alpha_1 y}{q}=r$

The real interest rate replaces the nominal rate in (2). It is no longer an expected rate and is now denoted by r rather than r^*. The term-structure relation is given by (5).

A. *Steady State*

In steady state $\dot{y}=0$, output equals spending which is given by

$$y=\frac{a}{b}q+\frac{1}{b}g$$

Output depends on the stock market and fiscal policy. (Recall that we do not allow prices to adjust and that, in this "steady state," output is demand determined, an assumption that we shall want to relax later.) From (2′) and (6′), if $\dot{q}=\dot{q}^*=0$:

$$q=\frac{\pi}{r}=\frac{\alpha_0+\alpha_1 y}{cy-h(m-p)}$$

The stock market is the ratio of steady-state profit to the steady-state interest rate. Both profit and interest are increasing functions of output: output increases profit directly; it also increases the transaction demand for money and the interest rate. The effect of output on the stock market is therefore ambiguous and two cases have to be considered:

Let $\bar{q}$ denote the steady-state value of q. Then if $(c\bar{q}-\alpha_1)>0$, then the interest rate effect will dominate and an increase in output will have the net effect of decreasing the stock market. For lack of a better term, this case will be called the *bad news* case. The other will be called the *good news* case.

The loci $\dot{y}=0$ and $\dot{q}^*=0$ are drawn in Figure 1; the steady state is characterized graphically in each of the two cases.[5]

B. *Dynamics*

In the absence of changes in current or future policies, the assumption of rational expectations implies that $\dot{q}^*=\dot{q}$. Thus the dynamic behavior of the economy is characterized by two differential equations in q

[5] In the good news case, the existence of an equilibrium where the ($\dot{q}=0$) locus intersects the *IS* from above follows from

$$\lim_{y\to\infty}\left.\frac{dq}{dy}\right|_{\dot{q}=0}=0$$

There might however be two equilibria. The other one however has both undesirable comparative statics and dynamic properties.

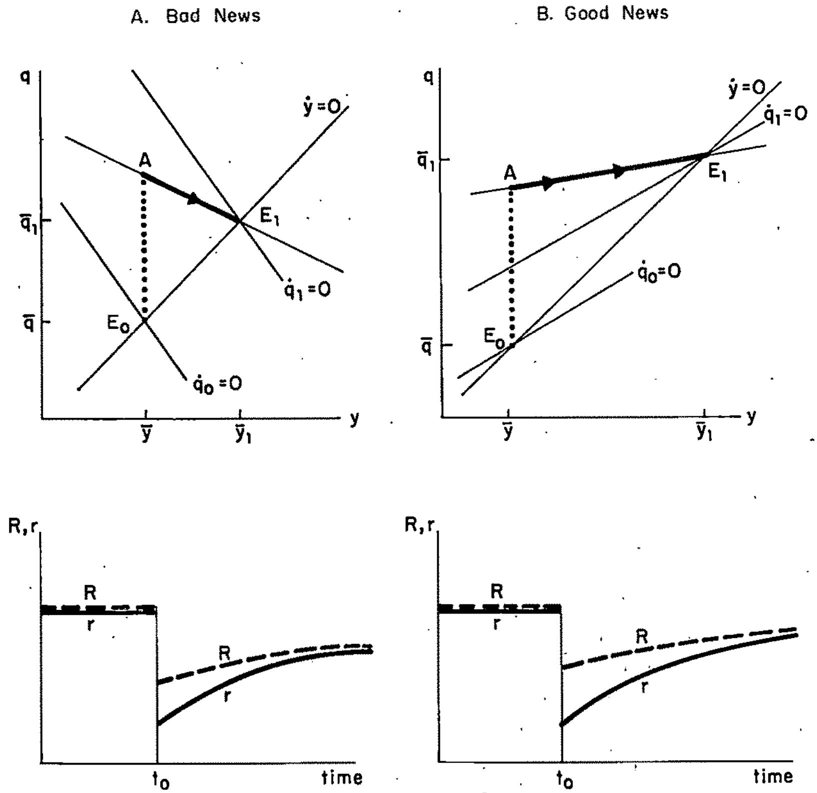


FIGURE 2. THE EFFECTS OF AN UNANTICIPATED MONETARY EXPANSION

and y. Trajectories corresponding to these equations of motion are drawn in Figure 1. In each case, the steady state is a saddle-point equilibrium. (Algebraic proofs and derivations are given in Appendix A.) Given the value of y which is given at any moment of time, there is a unique value of q which is such that the economy converges to its steady state. Following a standard if not entirely convincing practice,[6] I shall assume that q always adjusts so as to leave the economy on the stable path to equilibrium.

III. A Monetary Expansion under Fixed Prices

What are the effects of an increase in money? The "steady-state" effects are clear: output and the stock market are higher in the new equilibrium. The higher money stock lowers the real interest rate and thus the cost of capital. This lower cost leads to a higher stock market value, higher spending, higher output and profit. These comparative statics are very similar to the usual *IS-LM*. The dynamic adjustment is of more interest. To characterize it we need to distinguish between the case where the monetary expansion is unanticipated (i.e., announced and implemented simultaneously) and the case where it is anticipated (i.e., known for some time before its implementation).

A. *An Unanticipated Monetary Expansion*

The dynamic adjustment path is characterized in Figure 2 (for the system linearized around its initial steady state). It is drawn under the assumption that the economy is initially in steady state E_0. The stock market jumps to A and the economy converges to E_1 over time. The behavior of interest rates, which cannot be read off the phase diagram, is plotted below.

[6] See my 1979 paper for further discussion.

When the increase in money takes place, output is given and it is the short-term rate which falls to maintain portfolio balance. To understand why the stock market jumps, we can integrate forward the arbitrage equation (6) (taking into account the transversality condition imposed by the requirement that the economy converges). This gives q as the present discounted value of profits:

$$q_t = \int_t^{\infty} \pi^*(s) e^{-\int_t^s r^*(v)\,dv} ds$$

The jump is then easily understood by looking forward at the adjustment path: interest rates are anticipated to be lower than before and profits are anticipated to be higher. What happens to the long rate? Although the short-term rate initially falls, the expectation of increasing output leads to an expected increase in the demand for money and thus to an expected increase in the short rate: the long-term rate falls, but by less than the short; the term structure slopes upwards after the increase in money. Over time, production increases in response to spending. What happens to the stock market? As time passes, the initial low discount rates and low profits disappear from the integral and are "replaced" by higher discount rates and profits. In the bad news case, the discount rate effect dominates: the stock market initially overshoots its steady-state value and decreases thereafter. In this case consol prices and shares have a similar qualitative behavior. The opposite holds in the good news case: after their initial increase consol prices and shares move in opposite directions.

Note that the initial jump in the stock market is unanticipated, but that its movement afterwards is anticipated. This movement is in no way inconsistent either with rational expectations or the assumption of no excess return on shares. In this particular case, rational expectations for the stock market turn out to be regressive (see Appendix A). If $\bar{q}_1$ denotes the new steady-state value of q, we get

$$\dot{q}^* = \dot{q} = \mu(q - \bar{q}_1) \qquad \mu < 0$$

This result is however not very general and will not hold below.

B. *An Anticipated Monetary Expansion*

Suppose that the monetary expansion is announced at time t_0 to take place at time $t_1 > t_0$. The steady-state effects are the same as before, but the dynamic adjustment is different. It is characterized in Figure 3.

Technically, the adjustment path is uniquely determined by the following requirements. At time t_1, the economy, if it is to converge, must be somewhere on the stable arm of the postmonetary expansion system SS'. At time t_0, output $\bar{y}$ is given. Between t_0 and t_1 the system must satisfy the equations of motion on the premonetary expansion system. There is a unique trajectory which satisfies all these requirements; which one it is depends on the length of the period between t_0 and t_1. (Charles Wilson gives a more detailed explanation and the associated algebra in a model with a similar structure.) The curve $A'B'$ corresponds to a given period $(t_1 - t_0)$, $A''B''$ to a longer period.

The announcement of the monetary expansion is itself expansionary. The stock market jumps at time t_0 in anticipation of lower interest rates and higher profits after time t_1. This increases spending and output over time.

Between the announcement and the implementation, output increases. As the money stock is still constant, the short-term rate also increases. Because of anticipated lower short rates after t_1, however, the long-term rate declines. Thus the term structure "twists"; long and short rates moving in opposite directions. As the period of lower rates comes closer, the stock market increases. Whether or not it overshoots its steady state depends again on whether we are in the bad news or good news case.

At the time of the implementation, the short-term rate falls to maintain portfolio balance. *Little else happens*; the long-term rate and the stock market do not jump. If they did, this would imply anticipated infinite rates of capital gain or loss. After the implementation, the behavior of the economy is qualitatively similar to the case of an unanticipated increase.

Between t_0 and t_1, the movements in output, stock market, and interest rates happen

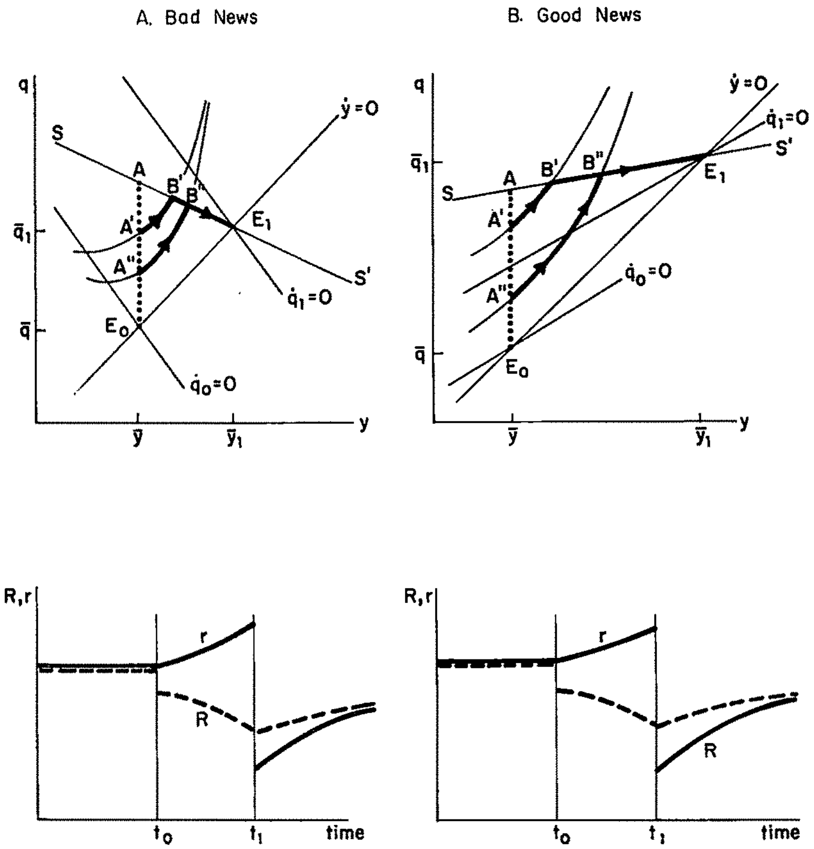


FIGURE 3. THE EFFECTS OF AN ANTICIPATED MONETARY EXPANSION

without apparent changes in policy. An outside observer, unaware of the announcement, may indeed conclude from an examination of the behavior of q and y over time, that an initial "speculative" boom in the stock market is the "cause" of the increase in output, and that the subsequent increase in money takes place to reduce the pressure on interest rates.

IV. A Fiscal Expansion under Fixed Prices

What are the effects of a fiscal expansion? The steady-state effects are again familiar: the fiscal expansion increases output, profit, and the interest rate. Hence, the effect on the stock market is ambiguous; the stock market decreases in the bad news case, increases in the other.

Turning to the dynamics, I shall directly consider the effect of an anticipated fiscal expansion, say, announced at t_0 to be implemented at time t_1. It is indeed probably the case that most changes in fiscal policy are known before they are implemented. (The case of an unanticipated change is just the limit of this case as t_1 tends to t_0.) The adjustment is characterized in Figure 4.

There are now important differences between the bad and good news cases. In the bad news case, the policy change is bad news for the stock market which falls at the time of announcement. The reason lies in the anticipated increase in the sequence of short-term rates after the policy is implemented. In this case, it more than compensates the anticipated increase in the sequence of profits. Between the announcement and the implementation, fiscal policy has a perverse effect on output: because of the decrease in the stock market, private spending decreases and public spending is

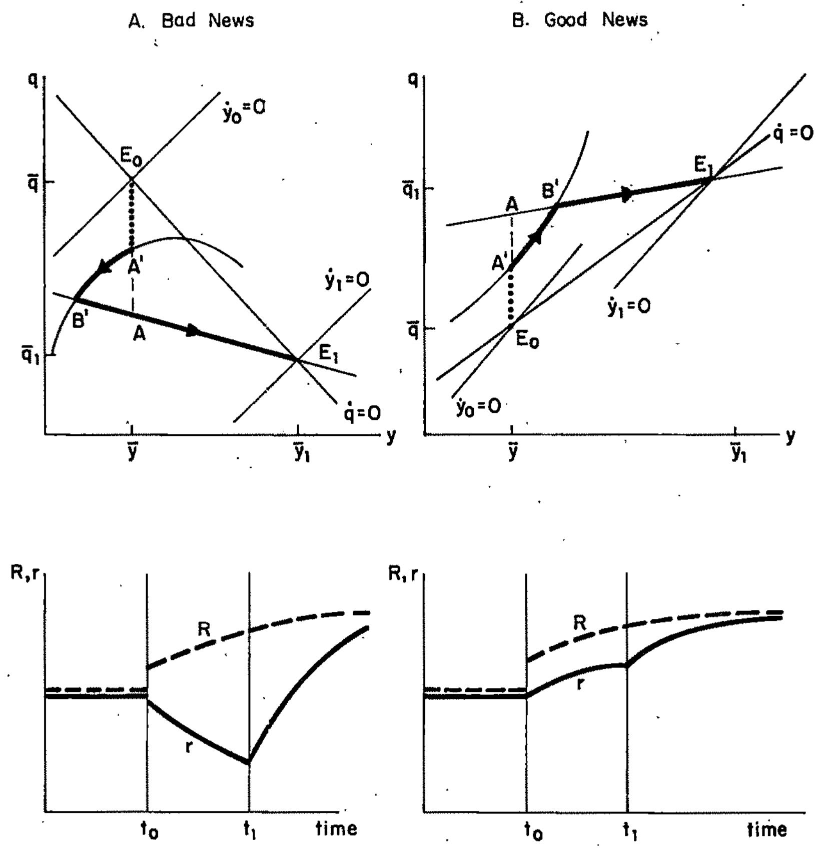


FIGURE 4. THE EFFECTS OF AN ANTICIPATED FISCAL EXPANSION

unchanged. Output therefore decreases until time t_1, and so does the short-term rate. The long-term rate, however, increases in anticipation of higher short rates in the future. At the announcement, the term structure is upward sloping and twists during the period between announcement and implementation.

At and after the implementation, public spending increases, and so does output over time. The short-term rate increases with output; the term structure remains positively sloped, its slope decreasing as the economy reaches its new steady state. The stock market and consol prices have, in this case, the same qualitative behavior.

There is no perverse effect of the announcement of a change in policy in the good news case. The anticipation of higher profits more than offsets the anticipation of higher rates and the stock market increases. This jump and subsequent increase in the stock market leads to an increase in output, and the long-term rate jumps in anticipation of higher short rates; the term structure is upward sloping. At the time of the implementation, the only noticeable effect is a larger rate of increase of output. An outside observer might again conclude that the policy change was not necessary, as the economy was already expanding. Note finally that, in this case, consol prices and the stock market move in opposite directions.

V. A Monetary Expansion under Flexible Prices

There is an uneasy feeling in characterizing the effects of nominal money over time assuming prices constant. I shall relax this assumption, at the cost of some additional complexity, and show how this affects the results obtained above. The discussion will be limited to characterizations of monetary policy.

If prices were perfectly flexible, changes in the level of money would be neutral, leaving output and the stock market unaffected. The dynamic adjustment of nominal variables could be characterized using this model (this would extend the analysis of Thomas Sargent and Neil Wallace) but would not be of considerable interest. We want instead to allow for movements in output, at least temporarily. The simplest price adjustment is probably

$$\text{(7)} \qquad \dot{p}=\dot{p}^*=\theta(\bar{p}-p) \qquad \theta>0$$

where $\bar{p}$ is the price level associated with full-employment output $\bar{y}$ and the level of nominal money $\bar{m}$. This adjustment process implies that money is neutral in the long run as prices adjust to their equilibrium value over time; it also assumes that the reason for the slow adjustment of prices is not irrational expectations, as $\dot{p}=\dot{p}^*$, but inertia or the existence of predetermined nominal contracts. The extension of equation (7) to include an unemployment term such as $(y-\bar{y})$ would make (7) look more like a Phillips curve. But it would make the analysis more cumbersome and not affect results substantially.

A. *Steady State and Dynamics*

From equations (1)–(7), steady-state output, interest rates, and the stock market are invariant to nominal money. Nominal money simply affects prices proportionately.

To understand how (7) affects the dynamics, consider an (unanticipated) increase in nominal money. When this expansion takes place, real balances are higher as prices cannot instantaneously adjust, decreasing the nominal interest rate. Prices are, however, now expected to increase and the expected rate of inflation decreases the real rate of interest given the nominal rate; this effect is usually referred to as the "Mundell effect." Both effects work in the same direction, decreasing the real rate. Over time, real balances decrease and the expected inflation becomes smaller; both effects work again in the same direction, now increasing the real rate. Algebraically, using (2), (3) and (7) gives

$$r^*=i-\dot{p}^*=cy-h(m-p)+\theta(p-\bar{p})$$

$$=(\theta+h)p+\psi$$

where ψ does not depend directly on p. A higher value of p leads to a higher r^* through the real balance effect (h) and the Mundell effect (θ).

Linearizing the system around its steady state gives

$$\text{(8)} \qquad \begin{bmatrix} \dot{q}^* \\ \dot{y} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} \bar{r} & c\bar{q}-\alpha_1 & (h+\theta)\bar{q} \\ \sigma a & -\sigma b & 0 \\ 0 & 0 & -\theta \end{bmatrix} \begin{bmatrix} q-\bar{q} \\ y-\bar{y} \\ p-\bar{p} \end{bmatrix}$$

This system has three roots. Because of its recursive structure, two of the roots are the same as in the fixed-price case, $\mu<0$ and $\xi>0$. The third is simply $-\theta$, the (negative of) the speed of adjustment of prices.

One root, ξ, is positive. Two variables y and p cannot "jump." Thus the initial conditions for (y,p) together with the requirement that the system converges to steady state determines a unique trajectory and, as in Sections II–IV, a unique value for q.

B. *An Unanticipated Monetary Expansion*

Consider an unanticipated monetary expansion dm at time $t_0=0$. If we assume the economy to be in steady state before the change, the behavior of output, prices and the stock market is given by (see Appendix B)

$$\text{(9)} \qquad q-\bar{q}=-\frac{\theta+h}{(\theta+\mu)(\theta+\xi)} \times\left[(\sigma b-\theta)e^{-\theta t}-(\sigma b+\mu)e^{\mu t}\right]\bar{q}\,dm$$

$$\text{(10)} \qquad y-\bar{y}=-\frac{\theta+h}{(\theta+\mu)(\theta+\xi)} \times\left[e^{-\theta t}-e^{\mu t}\right]\sigma a\bar{q}\,dm$$

$$\text{(11)} \qquad p-\bar{p}=-e^{-\theta t}dm$$

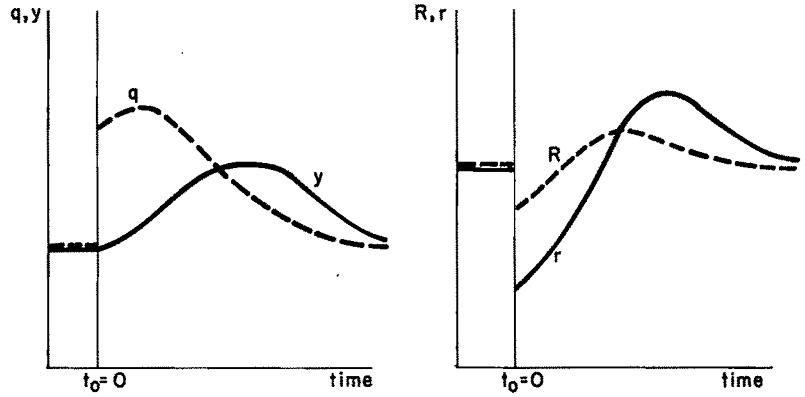


FIGURE 5. THE EFFECTS OF AN UNANTICIPATED MONETARY EXPANSION UNDER FLEXIBLE PRICES

(These relations hold only if $\theta \neq -\mu$. The Appendix gives the relations for $\theta = -\mu$.) If there was no Mundell effect (i.e., if, for example, money balances paid the expected rate of inflation), the relations defining q and y would have h rather than $(\theta+h)$ in the numerator. The roots θ, μ, ξ would be unaffected.

The first question is whether the analysis of Sections II–IV is the limit of the flexible price case when prices adjust extremely slowly, that is, as θ tends to zero. As shown in the Appendix, this is indeed the case. Thus the dynamics of Sections II–IV are "approximately correct" if prices adjust slowly.

The second question is how price flexibility affects the impact effect of monetary policy. Intuition suggests the elements of the answer: Assume that there is no Mundell effect; money balances pay the expected rate of inflation. Then the more flexible prices are, the faster the real money stock will return to its previous level. This in turn suggests a faster return of profits and real interest rates to their previous levels, thus a smaller initial jump in the stock market. This in turn implies a lower initial increase in spending and a lower rate of increase of output.

The countervailing effect of more flexible prices is through the Mundell effect. The more flexible prices are, the higher the initial rate of inflation and, *ceteris paribus*, the lower the real rate of interest. This lower initial sequence of real interest rates tends to increase the initial jump in the stock market, leading to a higher initial rate of increase in output.

Evaluating (9) at t = 0 gives

$$(q-\bar{q})|_{t=0} = \frac{\theta+h}{\theta+\xi}\bar{q}\,dm > 0$$

As θ, h, and ξ are positive, a monetary expansion always increases the stock market and spending. The effect of an increase in θ is indeed ambiguous. With no Mundell effect, the numerator would simply be h: in this case more flexible prices lead to a smaller impact effect. The Mundell effect works in the opposite direction; the net effect depends on $(h-\xi)$. The positive root ξ of the system does not depend on h. Thus there are no restrictions on the sign of $(h-\xi)$.

What can we deduce about the adjustment path from equations (9)–(11)? The monetary expansion leads to an expansion of output over time followed by a return to steady state; output remains above its equilibrium value during the period of adjustment. The real interest rate increases rapidly after its initial decline, overshooting its equilibrium value. This is due partly to lower real money balances, partly to higher transaction demand for money, and partly to lower expected inflation. As a result, the long-term real rate initially declines by less than the short rate; the term structure is initially upward sloping, flattening over time to become downward sloping. The sequences of initially increasing profits and interest rates followed by decreasing profits and interest rates have a complex effect on the stock market which may keep increasing after its initial jump and which may also decrease below its steady-state value during the adjustment process. These results are derived in Appendix B and summarized graphically in Figure 5.

VI. Summary and Extensions

This paper has shown the interaction between output and the stock market. The effect of a change either in current or anticipated policy is a discrete change in the stock market due to the change in the anticipated sequence of profits and real interest rates. This, in turn, together with the change in policy, affects spending and output over time, validating the initial anticipations of profits and interest rates. The stock market is not the "cause" of the increase in output, no more than the increase in output is the cause of the initial stock market change. They are both the results of changes in policy.

Whether policies are anticipated or not is important; the announcement itself will usually lead to a change in anticipated profits and discount rates, leading to a change in the stock market. Although in this case the change in the stock market and the resulting increase in output will precede the change in policy, they are still caused by it; the implementation of the policy may have little apparent effect. Under plausible assumptions, the announcement of an expansionary fiscal policy may have a perverse effect, decreasing output before the actual implementation of the policy.

The effect of more flexible prices is to decrease the overall effect of changes in nominal money. The effect on the initial impact of changes in money is, however, ambiguous: the smaller overall effect leads to a smaller change in the stock market, but the faster initial inflation may lead to lower real interest rates initially, and to a larger initial change in the stock market.

There are at least two logical extensions of this model. The first is a more detailed treatment of aggregate demand, allowing in particular a more detailed treatment of fiscal policy. This is done in a medium-sized econometric model (see my 1980 paper) and parallels a similar attempt by Ray Fair. The second is a more detailed treatment of aggregate supply: the implicit assumption of this paper is that the economy is in Keynesian unemployment, with no effects of the real wage. Allowing for an effect of the real wage and taking explicit account of capital accumulation are the next items on the agenda.

Appendix A: The Fixed-Price Case

The system composed of (1), (2′), and (6′) is linearized around its steady state:

$$\begin{bmatrix} \dot{q}^* \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \bar{r} & c\bar{q}-\alpha_1 \\ \sigma a & -\sigma b \end{bmatrix} \begin{bmatrix} q-\bar{q} \\ y-\bar{y} \end{bmatrix}$$

The assumption that the *IS* intersects the *LM* from below implies that the determinant of the above Jacobian is negative. The system has two roots of opposite sign, $\mu<0$ and $\xi>0$. The characteristic vector associated with μ, (x_1, x_2) is such that

$$\text{(A1)} \qquad \frac{x_1}{x_2} = \frac{c\bar{q}-\alpha_1}{\mu-\bar{r}} = \frac{\sigma b+\mu}{\sigma a}$$

The equations of motion along the stable arm are thus given by

$$\text{(A2)} \qquad q-\bar{q}=c_1(\sigma b+\mu)e^{\mu t}$$

$$y-\bar{y}=c_1\sigma a\, e^{\mu t}$$

where c_1 depends on the initial conditions for y. Thus, along the stable arm:

$$\text{(A3)} \qquad \frac{dq}{dy} \gtrless 0 \Leftrightarrow (\sigma b+\mu) \gtrless 0 \Leftrightarrow (c\bar{q}-\alpha_1) \lesseqgtr 0$$

An increase in nominal money, *dm*, increases steady state q and y. Denoting their new steady-state values by $\bar{q}_1$ and $\bar{y}_1$, we get

$$\text{(A4)} \qquad (\bar{q}_1-\bar{q})=-\frac{\sigma b\bar{q}h}{\mu\xi}dm$$

$$(\bar{y}_1-\bar{y})=-\frac{\sigma a\bar{q}h}{\mu\xi}dm$$

At the time of the increase, $t=0, y$ is fixed:

$$(y-\bar{y}_1)|_{t=0}=\bar{y}-\bar{y}_1\Rightarrow c_1=\frac{\bar{q}h}{\mu\xi}dm$$

This gives, replacing c_1 in (A2)

$$(q-\bar{q}_1)=\frac{\bar{q}h}{\mu\xi}(\sigma b+\mu)e^{\mu t}\,dm$$

Or, using (A4)

$$\text{(A5)} \quad (q-\bar{q})=\frac{\bar{q}h}{\mu\xi}((\sigma b+\mu)e^{\mu t}-\sigma b)\,dm$$

Appendix B: The Flexible Price Case

The characteristic polynomial of the matrix in (8) has three roots. Two are the same as in the fixed-price case, $\mu<0$ and $\xi>0$. The last is $-\theta<0$.

The vectors associated with the negative roots μ and $(-\theta)$ are, respectively (up to a factor of proportionality), if $\mu\neq-\theta$

$$\begin{bmatrix}\sigma b+\mu & \sigma a & 0\end{bmatrix}$$

$$\text{and} \quad \begin{bmatrix}\sigma b-\theta & \sigma a & \dfrac{(\theta+\xi)(\theta+\mu)}{(h+\theta)\bar{q}}\end{bmatrix}$$

The stable trajectory therefore satisfies

$$\text{(A6)} \quad q-\bar{q}=c_1(\sigma b-\theta)e^{-\theta t}+c_2(\sigma b+\mu)e^{\mu t}$$

$$\text{(A7)} \quad y-\bar{y}=c_1\sigma a\,e^{-\theta t}+c_2\sigma a\,e^{\mu t}$$

(A8)

$$p-\bar{p}=c_1(\theta+\xi)(\theta+\mu)(h+\theta)^{-1}\bar{q}^{-1}e^{-\theta t}$$

The variables c_1, c_2 are determined by initial conditions for y and p. An increase in dm leaves $\bar{y}$, $\bar{q}$ unchanged and increases $\bar{p}$ by dm. So after an unanticipated increase dm at $t_0=0$, $p-\bar{p}=-dm$ and $y-\bar{y}=0$. This determines uniquely c_1 and c_2 using (A7) and (A8) at $t_0=0$. Replacing these values of c_1 and c_2 in (A6) to (A8) gives equations (9)–(11) in the text. Note that if $\theta=0$, (9) reduces to equation (A5).

For the case where $\mu=-\theta$, L'Hospital's rule can be used on (9) and (10) to derive $(y-\bar{y})$ and $(q-\bar{q})$. This gives

$$q-\bar{q}=\bar{q}(\mu-h)(\mu-\xi)^{-1}$$
$$\times(1-t(\sigma b+\mu))e^{\mu t}\,dm$$

$$y-\bar{y}=\bar{q}\sigma a(\mu-h)(\mu-\xi)^{-1}t\,e^{\mu t}\,dm$$

We may now characterize the dynamic behavior of q. The initial jump in q at $t=0$ is given by

$$(q-\bar{q})|_{t=0}=\frac{\theta+h}{\theta+\xi}\bar{q}\,dm>0$$

As $(q-\bar{q})$ is the sum of two declining exponentials, it has for $t\geqslant0$ at most one interior maximum or minimum. If such an extremum exists, it happens at t^* given by

$$e^{-(\theta+\mu)t^*}=-\frac{\mu}{\theta}\frac{\sigma b+\mu}{\sigma b-\theta} \qquad \text{for } \theta\neq-\mu$$

$$t^*=-\sigma b\mu^{-1}(\sigma b+\mu)^{-1} \text{ for } \theta=-\mu$$

The initial rate of change of q is given by

$$\dot{q}|_{t=0}=-\frac{(\theta+h)}{(\theta+\mu)(\theta+\xi)}$$
$$\times[-\theta(\sigma b-\theta)-\mu(\sigma b+\mu)]\bar{q}\,dm$$
$$=\frac{\theta+h}{\theta+\xi}[\sigma b+\mu-\theta]\bar{q}\,dm$$

Thus, for $\dot{q}|_{t=0}>0$, a necessary condition is that $\sigma b+\mu>0$, or equivalently $c\bar{q}-\alpha_1<0$: the good news condition is necessary but not sufficient anymore.

The same analysis is easily done for y. The variable y increases initially, reaches its maximum for:

$$t^{**}=\frac{-ln(-\mu)+ln(\theta)}{\theta+\mu} \text{ for } \theta\neq-\mu$$

$$1/\theta \text{ for } \theta=\mu$$

It decreases to steady state after that. The behavior of the short-term real rate is directly obtained from

$$(r-\bar{r})=c(y-\bar{y})+(h=\theta)(p-\bar{p})$$

REFERENCES

O. Blanchard, "Backward and Forward Solutions for Economies with Rational Expectations," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 114–18.

_____, "The Monetary Mechanism in the Light of Rational Expectations," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago 1980.

R. C. Fair, "An Analysis of a Macro-Econometric Model with Rational Expectations in the Bond and the Stock Market," *Amer. Econ. Rev.*, Sept. 1979, *69*, 539–52.

L. Metzler, "Wealth, Savings and the Rate of Interest," *J. Polit. Econ.*, Apr. 1951, *59*, 93–116.

T. Sargent and N. Wallace, "The Stability of Models of Money and Growth with Perfect Foresight," *Econometrica*, Nov. 1973, *41*, 1043–48.

J. Tobin, "Monetary Policies and the Economy: The Transmission Mechanism," *Southern Econ. J.*, Jan. 1978, *44*, 421–31.

C. Wilson, "Anticipated Shocks and Exchange Rate Dynamics," *J. Polit. Econ.*, June 1979, *87*, 639–47.

Deregulation and Oligopolistic Price-Quality Rivalry

By JAMES H. VANDER WEIDE AND JULIE H. ZALKIND*

Domestic airlines, commercial banks, and motor trucking are oft-cited examples of industries in which price and the number of competitors are regulated, but firms compete freely on product quality. Other industries, such as commercial banks in limited-branching states, are more usefully thought of as being regulated on the quality dimension, but free to compete on price. Congress is considering several proposals to deregulate both types of industry. This paper presents a model of regulated, oligopolistic rivalry that can help evaluate congressional proposals for deregulation.

Deregulation of price-, quality-, and entry-regulated industries will be discussed in this paper within the framework of an explicit model of oligopolistic rivalry. The earlier work of Lawrence White, David Levhari and Yoram Peles, James Rosse, Elisha Pazner, and Rosse and John Panzar analyzed the effect of regulation on quality offerings in both monopolistic and perfectly competitive industries. These authors either said nothing directly about quality offerings in other industries, or indicated that oligopolistic results lie between those of monopoly and perfect competition. Panzar and Richard Schmalensee have shown that additional insight about the economic performance of oligopolistic markets such as airlines, banking, and trucking can be obtained from explicit models of oligopolistic behavior.

The present work differs from those of Panzar and Schmalensee in three important respects. First, we present a model of price variation in oligopolistic, *quality*-regulated markets such as commercial banking in limited-branching states, as well as a model of quality variation in price-regulated markets. Panzar and Schmalensee only consider models of the latter type.

Second, we extend the analysis of Panzar and Schmalensee to include situations of increasing or decreasing returns to scale in quality. Since Arthur DeVany finds evidence of increasing unit flight costs in the airline industry, and both George Benston and Frederick Bell and Neil Murphy find evidence of decreasing returns to scale in bank branches, we believe this extension significantly broadens the applicability of our results.

Finally, we explicitly analyze the economic effects of price, quality and/or entry deregulation of oligopolistic markets. Our analysis of deregulation is particularly relevant at a time when Congress is deregulating the airline, banking, and trucking industries, the CAB is allowing airline firms more freedom to lower fares and enter markets, and state legislatures are loosening restrictions on bank branches. The comparative static analyses of our predecessors do not address the economic effects of these changes directly.

The economic consequences of deregulation depend on a variety of factors, including the previously regulated variables, the choice of variables to deregulate, the appropriate type of individual firm cost and demand functions, and the relationship of the fixed values of the regulated variables to those which would obtain in an unregulated oligopoly. If entry into airline markets is deregulated, for instance, then the number of flights per carrier, the number of passengers per carrier and the load factor de-

*Associate professor, Graduate School of Business Administration, Duke University, and director, Mid-term Analysis Division, Energy Information Administration, U. S. Department of Energy, respectively. We would like to thank our colleagues Wesley Magat and David Peterson for their helpful critiques of an earlier draft. We would also like to thank Richard Arnould for suggesting the applicability of this model to safety regulation and David Zalkind for suggesting the means of incorporating safety and the derivation of the results for $\partial x/\partial \bar{p}$ in the advertising case.

crease, but total flights serving the market and total passengers in the market increase. If both price and the number of competitors are deregulated, however, the increase in market flights resulting from entry deregulation may be more than offset by a decrease in market flights resulting from price deregulation. Regulators should be aware of these dependencies when choosing a deregulation policy.

I. The Model

The oligopolistic firm of our model produces a product with both quantity and quality dimensions. Although the firm's revenue is explicitly based only on the quantity sold, the demand for its product depends on its own price and quality, as well as its competitor's price and quality levels.

The meaning of the product quantity, quality, and price dimensions depends on the market context. Within the context of airline regulation, quantity is often measured by the number of passengers, quality is measured by either the level of amenities per passenger or the number of flights serving each market, and price is measured by the ticket fare. Within the context of bank regulation, quantity might be measured by the number of customers served, quality by either the number of bank branches or the amount of free services per customer, and price by the interest offered on deposits.

The ensuing analysis depends on three simplifying assumptions: the assumption that the n firms in the market face identical demand and cost conditions; the assumption that the demand and cost functions take special forms; and the Cournot-Nash behavioral assumption that each firm views its rival's decisions as fixed. These assumptions reflect the economic character of price-quality rivalry in many oligopolistic markets. (Evidence in support of this assertion is cited in Schmalensee, fn. 3.)[1] They also permit analytic tractibility. Except for our generalization of their demand and cost functions, the above assumptions are identical to those chosen by Panzar and Schmalensee.

Let x_i be the quantity of firm i's product, q_i be the quality of firm i's product, p_i its price, and n the number of firms in the industry. Then the demand function is

$$(1)\quad x_i = G^0(q_i; q_1,\ldots,q_{i-1},q_{i+1},\ldots,q_n; p_i; p_1,\ldots,p_{i-1},p_{i+1},\ldots,p_n)$$

The identical firms assumption implies that G^0 is invariant with respect to permutations of its second through nth arguments and also with respect to permutations of its $n+2nd$ through $2n$th arguments. Thus, G^0 is the same for each firm. We assume that G^0 is homogeneous of degree α, $0 \leqslant \alpha < 1$, in its first n arguments and that G^0 is homogeneous of degree $-\gamma$, $0 \leqslant \gamma < 1$ in its final n arguments. The terms α and $-\gamma$ are the elasticity of firm i's quantity with respect to a simultaneous proportional change in all the firms' quality and in all the firms' price levels, respectively.

In the unregulated case, firm i's objective is to select q_i and p_i to maximize its profits

$$(2)\quad \pi_i(q_1,\ldots,q_n,p_1,\ldots,p_n) = p_i x_i - C(q_i, x_i)$$

where $C(q_i, x_i)$ is the (identical) cost function faced by all firms in the industry.

Under the Cournot and identical firms assumptions, firms make identical decisions. Thus, firm i can view its rival's choices as identical and equal to $\hat{q}$, $\hat{p}$, and it can view its own demand function (1) as

$$(3)\quad x = G(q,\hat{q},p,\hat{p},n) = G^0(q;\hat{q},\ldots,\hat{q};p;\hat{p},\ldots,\hat{p})$$

The function G is homogeneous of degree α in its first two arguments and of degree $-\gamma$ in its third and fourth arguments. Similarly, we can write the demand function as in (3) in the case where p or q is regulated.

The homogeneity of the demand function implies (as shown in the Appendix) that

[1] The evidence in Schmalensee only provides support for these assumptions in the quality rivalry case. The argument in favor of the Cournot-Nash behavioral assumption is less convincing when firms compete on price. However, the Cournot-Nash assumption permits analytical tractibility and the results are likely to be indicative of those in a more realistic, but less tractible, setting.

there exists $g(n)$ such that

$$(4)\qquad G(q,q,p,p,n)=q^{\alpha}p^{-\gamma}g(n)$$

Denote the partial derivative of $G(q,\hat{q},p,\hat{p},n)$ with respect to the first argument q as $G_1(q,\hat{q},p,\hat{p},n)$. Then when $\hat{q}=q$ and $\hat{p}=p$, there exists $k(n)$ such that

$$(5)\qquad G_1(q,q,p,p,n)=q^{\alpha-1}p^{-\gamma}k(n)$$

Similarly, there exists $\rho(n)$ such that

$$(6)\qquad G_3(q,q,p,p,n)=q^{\alpha}p^{-\gamma-1}\rho(n)$$

Economic considerations suggest that quantity demanded is positive, that demand for a firm's product decreases as additional firms enter the market, that a firm's demand increases as its quality increases and that a firm's demand decreases as its price increases. These considerations imply

$$(7)\qquad g(n)>0, g'(n)<0, k(n)>0,$$

and $\qquad \rho(n)<0$

The elasticities of demand with respect to changes in the firm's own price and quality decisions are

$$(8)\quad E(n)=-\frac{\partial x/\partial p}{x/p}=-\frac{G_3P}{G}=-\frac{\rho(n)}{g(n)}$$

$$(9)\quad B(n)=\frac{\partial x/\partial q}{x/q}=\frac{G_1q}{G}=\frac{k(n)}{g(n)}$$

respectively. Both $E(n)$ and $B(n)$ are positive under assumptions (7).

The following special case of demand function (1), which we call the *market share rivalry* demand function, sometimes provides additional insight

$$(10)\quad x_i=\frac{q_i^e}{\sum_{j=1}^{n} q_j^e}\left(\sum_{j=i}^{n} q_j\right)^{\alpha}\left[\frac{a}{p_i}-\frac{b(n-1)}{\sum_{j\neq i} p_j}\right]$$

where $\qquad 1>\frac{b}{a}>\frac{\gamma+1}{2}$

This demand function is a generalization, to permit price rivalry as well as quality rivalry, of the special market share scheduling rivalry demand function used by Joseph Yance, George Douglas and James Miller, and Schmalensee. In this demand function $(\Sigma q_j)^{\alpha}$ measures the effect of total market quality and the term $q_i^e/\Sigma q_j^e$ measures the effect of changes in firm i's market share of quality. Quantity is negatively related to the firm's own price and positively related to the average price charged by other firms. For this demand function, we show in the Appendix that

$$g(n)=n^{\alpha-1}(a-b)^{\gamma}$$

$$k(n)=n^{\alpha-2}(ne+\alpha-e)(a-b)^{\gamma}$$

$$\rho(n)=-n^{\alpha-1}\left(\gamma a(a-b)^{\gamma-1}\right)$$

Then $E'(n)=0$ and if $\alpha<1<e$, $B'(n)>0$ and $\rho'(n)<0$.[2]

The demand function (1) permits the study of different types of regulation, including price regulation, quality regulation, and entry regulation. For example, the government may regulate airline price and restrict entry, leaving the quality choice to the firm. In contrast, the government may regulate both bank entry and a quality variable such as the number of branches per bank, and the firm is free to set price.

We will use two different types of cost functions, depending on the interpretation of the quality variable. The first cost function is appropriate to situations, such as branch banking and airline flight schedules, where the costs of additional quality are independent of the number of customers served.[3] This cost function is[4]

$$(11)\qquad C(x,q)=dx+cq^{\beta}$$

[2] Douglas and Miller show that $\alpha<1<e$ is characteristic of the airline industry.

[3] Throughout the remainder of our analysis, we assume that the quantity variable x also measures the number of customers served. We noted earlier that this is appropriate for both the airline and banking industries.

[4] Panzar and Schmalensee use the version of this cost function where $\beta=1$.

where c, d and β are nonnegative. The variable β measures the economies of scale with respect to quality in this case.

The second type of cost function is appropriate to situations, such as automobile safety, where quality attaches to each unit of output. For this case, we use the cost function[5]

$$(12) \qquad C(x,q)=axq^{\delta}+dx$$

where a, δ, and d are nonnegative. The variable d is the cost of providing the basic unit of output, such as an automobile, to each customer and aq^{δ} represents the per customer cost of providing the quality of service q. The variable δ measures economies of scale in quality for this second case.

Although it is possible to combine the two types of cost functions into a single, more general form, the issues become considerably more complex when this is done. We believe our results are indicative of results that would obtain in a more complex setting.

II. Deregulation of Price and Entry when Quality is Independent of Output

Firms often compete on the basis of product quality in industries where price is fixed by regulatory control. The basic model can be adapted to such situations simply by fixing price at some level, say $\bar{p}$. The objective in this case is to choose q to maximize

$$(13) \qquad \pi(q)=\bar{p}.x-C(x,q)$$

If quality is independent of the number of customers served, we can rewrite equation (13) as

$$\pi(q)=(\bar{p}-d)G(q,\hat{q},\bar{p},\bar{p},n)-cq^{\beta}$$

Each firm maximizes $\pi(q)$ under the assumption that competitors' decisions are invariant with respect to changes in q. The initial first-order condition is

$$(\bar{p}-d)G_1(q,\hat{q},\bar{p},\bar{p},n)-c\beta q^{\beta-1}=0$$

Since all firms are identical, under the Cournot assumption, they will all produce the same quality, i.e., $\hat{q}=q$. In equilibrium then, we can use equation (5) to rewrite the first-order condition as[6]

$$(14) \qquad (\bar{p}-d)\bar{p}^{-\gamma}k(n)-c\beta q^{\beta-\alpha}=0$$

A typical firm's equilibrium reaction to a change in the regulated price p, and number of firms n, is found through implicit differentiation of (14). Table 1 displays the comparative static results developed in the Appendix. The variables of interest include quality q and quantity x per firm, market output of quantity nx, quality nq, quality per customer q/x, and firm profits π.

A. *Banking*

We can use results of Table 1 to evaluate some current congressional proposals to deregulate the commercial banking industry. (In our interpretation of these results, we let the quality variable q measure branches per bank, the quantity variable x measure the number of customers per bank, and the variable n measure the number of banks in the market.) Most proposals call for an increase in the number of competitors and removal of all restrictions on the interest rates banks pay on demand and time deposits. In the context of our model, we evaluate these proposals by considering the effects of 1) a decrease in price $\bar{p}$ (a decrease in price is equivalent to an increase in interest rates), 2) an increase in the number of competitors n, and 3) a simultaneous decrease in price and increase in the number of competitors.

The first line of Table 1 describes both the firm and industry response to a decrease in price, holding the number of firms fixed.

[5] Levhari and Peles treat a more general form of this cost function. White treats the case where $\delta=1$ and Pazner treats a more general version of this cost function but with $d=0$. None of these analyses, however, apply to oligopolistic industries.

[6] The solution to this equation exists, is unique among symmetric solutions and satisfies second-order and entry restrictions as long as $(p-d)>0$, $G_{11}<0$, and $\beta>1$. Second-order conditions may also be satisfied for some values of $\beta<1$.

TABLE 1—COMPARATIVE STATIC EFFECTS OF CHANGES IN THE REGULATED VARIABLES p AND n WHEN QUALITY IS INDEPENDENT OF OUTPUT

Regulated			Endogenous Variables			
Variables	q	x	nx	q/x	π	$Q=nq$
p	Pos	$\frac{\alpha}{\beta\gamma}-\frac{p-d}{p}$	$\frac{\partial x}{\partial p}$	Pos	Pos	Pos
n	$k'(n)$[b]	Neg	Pos	Pos[a]	Neg	Pos

Note: Definition of variables: p=output price; n=number of firms in industry; q=amount of quality, for example, branches or flights; x=quantity of output; π=profit; α=elasticity of firm i's quantity with respect to simultaneous proportional change in all the firms' quality levels; $-\gamma$=elasticity of firm i's quantity with respect to simultaneous proportional change in all the firms' price levels.

[a] True for market share rivalry demand function.

[b] In the case of the market share rivalry demand function, $k'(n)$ is negative for large n.

Branches per bank q, and total branches in the market nq, both decrease. This result is consistent with the results for a competitive firm found by White; namely, that when firms in a regulated competitive industry cannot compete on price, they compete away their excess profits through the quality variable. Since $d\pi/d\bar{p}>0$ in the oligopoly case, however, firms do not decrease quality so much that they compete away profits entirely.

A surprising result of our analysis is that the sign of $\partial x/\partial\bar{p}$ is indeterminate. We believe that $\partial x/\partial\bar{p}<0$ in most cases. In some cases, however, a decrease in the regulated price causes the bank to decrease its branches so much that demand actually decreases. This phenomenon occurs when $\alpha/\beta\gamma>(\bar{p}-d)/\bar{p}>1$. In this case, the market reaction to matched price changes measured by α is significantly greater than the product of the market reaction to matched quality changes measured by γ and the measure of economies of scale, β. Total industry output changes in the same direction as firm output.

The second line of Table 1 shows the responses to a change in the number of competitors, holding the regulated price fixed. The assumption that the banks entering (or leaving) the market are identical to the banks already in the market is implicit in the derivation of these results. An increase in n leads to a decrease in output per bank, but an increase in total output and branches in the market. Individual bank profits decline.

Policymakers could decide, of course, to allow free entry at the same time that they deregulate price. The effect of a simultaneous change in $\bar{p}$ and n can be found through analysis of the total differential:

$$dz=\frac{\partial z}{\partial\bar{p}}d\bar{p}+\frac{\partial z}{\partial n}dn$$

where z represents the variable of interest. For example,

$$dq=\frac{\partial q}{\partial\bar{p}}d\bar{p}+\frac{\partial q}{\partial n}dn \tag{15}$$

measures the effect of deregulation on the individual firm's quality choice. We know from Table 1 that $\partial\pi/\partial\bar{p}>0$ and $\partial\pi/\partial n<0$. Since simultaneous deregulation of both price and entry leads to lower p and higher n, deregulation will cause a decrease in individual bank profits. The effect of deregulation on total branches and output is indeterminate without empirical estimation of the demand and cost parameters.

B. *Airlines*

Recent models of the airline industry (see, for example, Douglas and Miller, Schmalensee, and Panzar) have assumed that regulatory authorities fix fares and the number of competitors, but allow firms to compete on their flight schedules. Both Congress and the CAB have deregulated the domestic airline industry. Most observers agree that deregulation will produce a decrease in price and/or an increase in the number of competitors. If this observation is correct, the remaining consequences of deregulation can be predicted from the results displayed in Table 1. For the purposes of this analysis, the quality variable q is the number of flights per firm, the quantity variable x is the number of customers served by each firm, and the variable n is the number of airline firms in the market.

If price alone is deregulated, then deregulation will lead 'to a lower fare, fewer total flights, and lower industry profits.[7] The load factor will increase as well. More customers will be served if price effects significantly dominate quality effects. The social benefit of such a change is ambiguous, since customers pay lower fares at the possible expense of a reduction in flights. In small markets in which the regulated price may be below the equilibrium unregulated fare, these effects are reversed.

An alternative deregulation policy is to allow free entry into a given airline market, but retain control over fares. Adoption of this alternative would lead to a decrease in the number of flights per carrier (in the case of the demand function (10)); a decrease in customers served per carrier but an increase in total customers served by the market. If the special demand function (10) applies, there will be a decrease in the load factor and firm profits, and an increase in total flights in the market.

Finally, a policy of total deregulation of the large airline market will cause a decrease in fares, a decrease in the number of flights per market, and a decrease in firm profits.[8]

What alternative should the deregulators adopt? The answer depends on the social welfare weights attached to each price-quantity-quality combination. By indicating the directions of the economic effects of deregulation, as well as the parameters which need measurement to estimate the magnitude of these economic effects, however, our analysis should help clarify the debate.

III. Price and Entry Deregulation when Quality Depends on Output

We use the cost function (12) to study the effects of deregulation when the quality variable of interest is attached to each unit of output sold, as in the airline meals case studied by White. In this case, profits are

$$\pi(q)=(\bar{p}-aq^{\delta}-d)G(q,\hat{q},\bar{p},\bar{p},n)$$

The Cournot equilibrium first-order condition for profit maximization is[9]

$$(\bar{p}-d)\frac{B(n)}{[1+B(n)]}-aq^{\delta}=0$$

where $B(n)$, the elasticity of demand with respect to matched changes in quality, is nonnegative.

The interpretation of quality in this case differs from the interpretation in the earlier case. Since quality in this case is inextricably attached to each unit of output sold, the variable q now measures quality per unit, and total firm quality offerings is measured by qx; that is, by quality per unit times the number of units sold. Similarly, the variable nqx measures total market quality.

Table 2 shows a comparative static analysis of the firm's reaction to changes in the regulatory variables p and n. The differences in the meaning of the quality variable lead to a number of differences between these results and those applicable to the previous cost function. First, the reactions of the quantity variables x and nx to changes in price restrictions are negative, and the reactions of excess profits π are positive only when $\gamma>\alpha$ and $\gamma\geqslant 1$. The reaction of the firm's and the market's total output of quality qx and nqx to changes in price is only determinate and positive when there are economies of scale in quality, i.e., $\delta\leqslant 1$.

[7] Firms will earn lower profits after deregulation if they are earning a positive economic profit prior to deregulation. If firms are not making positive economic profits prior to deregulation, then deregulation leads to higher profits, as shown in the Appendix. Casual empiricism suggest that this may be the case on many airline routes currently being deregulated.

[8] See fn. 7.

[9] The solution to this equation exists, is unique among symmetric solutions, and satisfies second-order and entry restrictions as long as $p-af^{\delta}-d>0$, $G_{11}<0$, and $\delta\geqslant 0$. Second-order conditions may be satisfied if $\delta<1$.

TABLE 2—COMPARATIVE STATIC EFFECTS OF CHANGES IN THE REGULATED VARIABLES p AND n WHEN QUALITY IS TIED TO OUTPUT

Regulated	Endogenous Variables					
Variables	q	x	nx	qx		π
p	Pos	$\frac{\alpha}{\gamma\delta}-\frac{p}{p-d}$	$\frac{\partial x}{\partial p}$	$\frac{\alpha+1}{\delta}-\frac{p-d}{p}$	$\frac{\partial qx}{\partial p}$	Pos
n	Pos[a]	$\frac{\alpha}{\delta}+\frac{g'}{g}\frac{B(1+B)}{B'}$	Pos[a]	$\frac{\alpha+1}{\delta}+\frac{g'}{g}\frac{B(1+B)}{B}$	Pos[a]	$\frac{\alpha}{\gamma}-1+\frac{g'}{g}\frac{B(1+B)}{B'}$[b]

Note: See Table 1; q=amount of quality per unit of output, δ=measure of economies of scale in quality, B=elasticity of demand with respect to changes in the firms' own price decision, g=a function reflecting the impact of the number of firms n on the quantity demanded of firm i's product.

[a] See Table 1.

[b] Negative for $\alpha<\gamma$.

Finally, in this case $\partial\pi/\partial n$ could be positive if α is much greater than γ. If the further assumptions that $\delta=1$ and $\gamma>\alpha$ are made, then Table 2 shows that price deregulation leads to more customers but fewer services per customer.

The market share rivalry demand function permits analysis of the effects of entry deregulation. As the number of firms in the market increase, services per customer and per market increase.

IV. Quality-Regulated Oligopolies: Branch Banking and Safety

Quality regulation characterizes some industries better than price regulation. Examples of these industries include banking in limited-branching states and safety-regulated industries such as automobiles. The economics of quality regulation can be analyzed within the context of our model by treating the firm's quality variable as fixed. Then, the firm's objective is to choose p to maximize

$$\pi(p)=p\cdot x-C(x,\bar{q}) \tag{16}$$

We use (16) with the appropriate cost function to analyze cases such as branching where quality is independent of the firm's output and cases, such as automobile safety, where quality and output are inextricably tied together.

A. *Branch Banking*

Many states limit the number of branches that may be operated by any one bank and/or restrict entry into banking markets. Our model permits analysis of the economic effects of branch and/or entry deregulation. We treat the number of branches as an advertising-type quality variable measuring "convenience" and assume that banks are free to compete on price.

If the cost of establishing a branch is independent of the number of customers served, then (16) can be written as

$$\pi(p)=(p-d)G(\bar{q},\bar{q},p,\hat{p},n)-c\bar{q}^{\beta}$$

and the profit-maximizing values of p is one which satisfies the first-order condition:[10]

$$p=\frac{dE(n)}{E(n)-1}$$

a solution which only makes economic sense when $E(n)>1$.

Table 3 presents comparative static results useful for evaluation of branch and entry deregulation. We assume that branches per firm q and the number of firms n are less than their equilibrium unregulated values. The first row of Table 3 indicates the effect of deregulating the number of branches per firm, holding the number of firms constant. The second row indicates the effect of allowing free entry, while continuing to regulate branches.

[10] A solution exists, is unique among symmetric solutions, and satisfies positive profit requirements for entry if $(p-d)>0$ and $G_{33}<0$.

TABLE 3—COMPARATIVE STATIC EFFECTS OF CHANGES IN THE REGULATED VARIABLES q AND n WHEN QUALITY COST IS INDEPENDENT OF OUTPUT

Regulated		Endogenous Variables			
Variables	p	x	nx	x/q	π
q	0	Pos	Pos	Neg	Indet[b]
n	0[a]	Neg[a]	Pos[a]	Neg[a]	Neg[a]

Note: See Table 1; q = the amount of quality, for example, the number of branches.
[a] See Table 1.
[b] Positive for $\beta < 1$.

TABLE 4—COMPARATIVE STATIC EFFECTS OF CHANGES IN THE LEVEL OF SAFETY WHEN SAFETY IS INDEPENDENT OF UNITS SOLD

Regulated		Endogenous Variables			
Variables	p	x	nx	q/x	π
q	0	0	0	Pos	Neg
n	0[a]	Neg[a]	0	Pos[a]	Neg[a]

Note: See Table 1; q = amount of safety when safety is independent of output.
[a] See Table 1.

TABLE 5—COMPARATIVE STATIC EFFECTS OF CHANGES IN THE LEVEL OF SAFETY WHEN SAFETY IS ATTACHED TO EACH UNIT SOLD

Regulated		Endogenous Variables		
Variables	p	x	nx	π
q'	Pos	Neg	Neg	Neg
n	Neg	Neg[a]	0[a]	Neg[a]

Note: See Table 1; q' = amount of safety when safety depends on output.
[a] See Table 1.

Deregulation of branching leads to an increase in both bank and industry output. Since the number of branches affects marginal revenue and marginal cost proportionately, however, deregulation has no effect on equilibrium price. Because consumers can receive both greater convenience and quantity at the same price, their welfare is improved by free branching. Consumer welfare always comes at the expense of the firm when there are diseconomies of scale ($\beta > 1$).

If the market share rivalry demand function applies, deregulation of entry leads to an increase in market output and no change in price. Thus, consumer welfare is improved by this form of deregulation as well. Since individual firm profits decline under free entry, improvements in consumer welfare come at the expense of the firm.

B. *Safety Regulation*

Safety regulation is a special case of quality regulation in which quality has little or no effect on demand. This phenomenon is especially apparent in the automobile industry where manufacturers frequently insist that safety has no effect on passenger demand for automobiles. It is also apparent in the airline industry where passengers do not possess the knowledge to select a carrier on the basis of safety.

There are two types of safety: one that attaches to each unit sold, as in automobile safety, and another which is independent of the number of customers, as in airline safety. Tables 4 and 5, derived under the assumption that $\alpha = 0$, show the comparative static effects of safety regulation for the cost functions corresponding to these two cases. The contrast is striking. Since increased safety standards have no effect on the firms' price and output decisions in the case where safety is independent of the number of customers served, consumer welfare is unambiguously improved, but firm profits decline. In the case where safety attaches to each unit sold, however, equilibrium market price increases and quantity decreases when safety standards are raised. Thus, consumers share the cost burden of higher safety with firms.

V. Conclusion

We have developed an explicit model showing the effects of deregulation of an oligopolistic industry. The oligopoly form of the model permits explicit analysis of entry as well as price and/or quality deregulation. Our model pertains to two interpretations of quality and also allows for nonconstant returns to scale in quality. The direction and magnitude of the effects of deregulation depends on which variables were previously regulated and which are deregulated. These effects were determined in many cases without resort to statistical estimation of model parameters. The results of this paper should aid in the evaluation of recent proposals to deregulate important sectors of the American economy.

Appendix

A: The Demand Function

LEMMA: *$G_1(q, \hat{q}, p, \hat{p}, n)$ is homogeneous of degree $\alpha-1$ in the first two arguments.*

PROOF:

Differentiating G, gives

(A1)

$$\frac{\partial G(\lambda q, \lambda \hat{q}, p, \hat{p}, n)}{\partial q} = \lambda G_1(\lambda q, \lambda \hat{q}, p, \hat{p}, n)$$

and by definition

(A2)

$$\frac{\partial G(\lambda q, \lambda \hat{q}, p, \hat{p}, n)}{\lambda q} = \frac{\partial \lambda^{\alpha} G(q, \hat{q}, p, \hat{p}, n)}{\partial q}$$

$$= \lambda^{\alpha} G_1(q, \hat{q}, p, \hat{p}, n)$$

Combining (A1) and (A2) gives

$$G_1(\lambda q, \lambda \hat{q}, p, \hat{p}, n) = \lambda^{\alpha-1} G_1(q, \hat{q}, p, \hat{p}, n)$$

LEMMA: *There exists $g(n)$ such that $G(q, q, p, p, n) = q^{\alpha} p^{-\gamma} g(n)$.*

PROOF:

$G(q, q, p, p, n) = q^{\alpha} p^{-\gamma}$ $G(1, 1, 1, 1, n)$ from the assumed homogeneity. Let $g(n) = G(1,1,1,1,n)$.

LEMMA: *For the market share rivalry demand function*

$$x_i = \frac{q_i^e}{\sum_{j=1}^{n} q_j^e} \left(\sum_{j=1}^{n} q_j \right)^{\alpha} \left[\frac{a}{p_i} - \frac{b(n-1)}{\sum_{j \neq i} p_j} \right]^{\gamma}$$

A. $g(n) = n^{\alpha-1}(a-b)^{\gamma}$
B. $k(n) = n^{\alpha-2}(ne+\alpha-e)(a-b)^{\gamma}$
C. $\rho(n) = -n^{\alpha-1}\gamma a(a-b)^{\gamma-1}$

PROOF:

A. Let $q_i = q, p_i = p$ for $i = 1, \ldots, n$. Then

$$x_i = \frac{q^e}{nq^e}(nq)^{\alpha}\left(\frac{a}{p} - \frac{b(n-1)}{(n-1)p}\right)^{\gamma}$$

$$= q^{\alpha} p^{-\gamma} n^{\alpha-1}(a-b)^{\gamma}$$

B. Differentiate x_i with respect to q_i to get

$$\frac{\partial x_i}{\partial q_i} = \left(\sum q_j^e\right)\left[e q_i^{e-1}\left(\sum q_j\right)^{\alpha} + \alpha\left(\sum q_j\right)^{\alpha-1} q_i^e \right]$$

$$-q_i^e\left(\sum q_j\right)\alpha e q_i^{e-1} + \left(\sum q_j^e\right)^2$$

$$x \left[\frac{a}{p_i} - \frac{b(n-1)}{\sum_{j \neq i} p_j} \right]^{\gamma}$$

Set $q_i = q$ and $p_i = p$ for $i = 1, \ldots, n$. Then

$$\frac{\partial x_i}{\partial q_i} = nq^e\left[e q^{e-1}(nq)^{\alpha} + \alpha(nq)^{\alpha-1} q^e \right]$$

$$-q^e(nq)^{\alpha} e q^{e-1} +$$

$$\div(nq^e)^2 x\left(\frac{a-b}{p}\right)^{\gamma}$$

$$=n^{\alpha-1}[en+\alpha-1]q^{\alpha-1}p^{-\gamma}(a-b)^{\gamma}$$

C. The proof of C is similar to the proof of B.

B: Comparative Statics

The comparative static results shown in Table 1 are derived in this Appendix. (The results in Tables 2–5 are derived in much the same way as those in Table 1; the interested reader may obtain these derivations from the authors.)

If p is regulated and the cost of quality is independent of output, then the firm's objective is to choose q to maximize

$$\text{(A3)}\quad \pi(q)=(\bar{p}-d)G(q,\hat{q},\bar{p},\bar{p},n)-cq^{\beta}$$

and the first-order condition is

$$(\bar{p}-d)G_1(q,\hat{q},\bar{p},\bar{p},n)-c\beta q^{\beta-1}=0$$

Substituting for G_1, setting $\hat{q}=q$ and rearranging yields

$$\text{(A4)}\quad (\bar{p}-d)\bar{p}^{-\gamma}k(n)-c\beta q^{\beta-\alpha}=0$$

To determine the effect of changes in the regulatory variable n, implicitly differentiate (A4) and substitute (A4) into the result to obtain

(A5)

$$\frac{\partial q}{\partial n}=\frac{(\bar{p}-d)\bar{p}^{-\gamma}k'(n)}{c\beta(\beta-\alpha)q^{\beta-\alpha-1}}=\frac{q}{\beta-\alpha}\cdot\frac{k'(n)}{k(n)}$$

Equation (A4) is used in a similar manner to obtain

$$\text{(A6)}\quad \frac{\partial q}{\partial \bar{p}}=\frac{k(n)\bar{p}^{-\gamma-1}[\bar{p}(1-\gamma)+d\gamma]}{c\beta(\beta-\alpha)q^{\beta-\alpha-1}}$$

$$=\frac{q}{\bar{p}}\cdot\frac{[\bar{p}(1-\gamma)+d\gamma]}{(\bar{p}-d)(\beta-\alpha)}$$

If profits are positive, then $\bar{p}>d$. Under the assumptions that $\beta>\alpha$ and $k'(n)<0$ (true in the special case) we then have $\partial q/\partial n<0$. Furthermore, since $\gamma<1$, when $\beta>\alpha$, $\partial q/\partial\bar{p}>0$. In equilibrium and from (A4)

$$\text{(A7)}\quad x=q^{\alpha}\bar{p}^{-\gamma}g(n)$$

$$=\left[\frac{(\bar{p}-d)k(n)\bar{p}^{-\gamma}}{c\beta}\right]^{\frac{\alpha}{\beta-\alpha}}\bar{p}^{-\gamma}g(n)$$

Differentiating *ln* x from (7) with respect to $\bar{p}$ we find that

$$\text{(A8)}\quad \operatorname{sgn}\frac{\partial x}{\partial\bar{p}}=\operatorname{sgn}\left(\frac{\alpha}{\beta\gamma}-\frac{\bar{p}-d}{\bar{p}}\right)$$

Differentiating (A7) with respect to n yields

$$\text{(A9)}\quad \frac{\partial x}{\partial n}=q^{\alpha-1}\bar{p}^{-\gamma}\left[\alpha g(n)\frac{\partial q}{\partial n}+qg'(n)\right]$$

We know that $\partial q/\partial n<0$, as derived above, and from the assumption that $g'(n)<0$ that $\partial q/\partial n<0$.

Let $Q=nq$ be total quality produced by the market. Then

$$\frac{\partial Q}{\partial\bar{p}}=n\frac{\partial q}{\partial\bar{p}}>0$$

and
$$\frac{\partial Q}{\partial n}=n\frac{\partial q}{\partial\bar{p}}+q\gtrless 0$$

Let $X=nx$ be market output. Then

$$\frac{\partial X}{\partial n}=n\frac{\partial q}{\partial n}+q\gtrless 0$$

Let $x/q=q^{\alpha-1}\bar{p}^{-\gamma}g(n)$ be customers per unit of quality. Then using (A6)

$$\frac{\partial(x/q)}{\partial\bar{p}}=q^{\alpha-2}\bar{p}^{-\gamma-1}g(n)\cdot$$

$$\left[(\alpha-1)\bar{p}\frac{\partial q}{\partial\bar{p}}-\gamma q\right]<0$$

and
$$\frac{\partial(x/q)}{\partial n}=q^{\alpha-2}\bar{p}^{-\gamma}$$

$$\left[(\alpha-1)g(n)\frac{\partial q}{\partial n}+qg'(n)\right]\gtrless 0$$

Therefore $\dfrac{\partial q/x}{\partial\bar{p}}>0$.

Equilibrium profits are found by substituting the cost function into (A3), and using (A4) and the definition of $B(n)$:

$$\text{(A10)} \quad \pi=(\bar{p}-d)\bar{p}^{-\gamma}q^{\alpha}g(n)-cq^{\beta}$$

$$=cq^{\beta}\left[\frac{\beta}{B(n)}-1\right]$$

When economic profits are positive, $\beta/B(n)>1$. Then, differentiating (A10) with respect to $\bar{p}$ and using (A6) we obtain

$$\frac{\partial\pi}{\partial\bar{p}}=c\beta q^{\beta-1}\left(\frac{\beta}{B(n)}-1\right)\frac{\partial q}{\partial\bar{p}}$$

which is positive when economic profits are positive. Differentiating (A9) with respect to n yields

$$\frac{\partial\pi}{\partial n}=c\beta q^{\beta-1}\left[\left(\frac{\beta}{B(n)}-1\right)\frac{\partial q}{\partial n}-\frac{qB'(n)}{B(n)^2}\right]$$

Since $(\beta/B(n)-1)>0$, $\partial q/\partial n<0$ and $B(n)>0$, then $\partial\pi/\partial n<0$, if $B'(n)>0$, which is true for the special case of our demand function.

REFERENCES

F. W. Bell and N. B. Murphy, "The Impact of Market Structure on the Price of a Commercial Bank Service," *Rev. Econ. Statist.*, May 1969, *51*, 210–13.

G. J. Benston, "Economies of Scale and Marginal Costs in Banking Operations," *Nat. Bank. Rev.*, June 1965, *2*, 507–49.

A. S. DeVany, "The Effect of Price and Entry Regulation On Airline Output, Capacity and Efficiency," *Bell J. Econ.*, Spring 1975, *6*, 327–45.

G. W. Douglas and J. C. Miller, "Quality Competition, Industry Equilibrium, and Efficiency in the Price-Constrained Airline Market," *Amer. Econ. Rev.*, Sept. 1974, *64*, 657–69.

D. Levhari and Y. Peles, "Market Structure, Quality and Durability," *Bell J. Econ.*, Spring 1973, *4*, 235–48.

J. C. Panzar, "Regulation, Service Quality, and Market Performance: A Model of Airline Rivalry," memo. no. 184, Center Res. Econ. Growth, Stanford Univ., Jan. 1975.

E. A. Pazner, "Quality Choice and Monopoly Regulation," in Richard E. Caves and Marc J. Roberts, *Regulating the Product: Quality and Variety*, Cambridge, Mass. 1975, 3–16.

J. N. Rosse, "Product Quality and Regulatory Constraints," memo. no. 137, Center Res. Econ. Growth, Stanford Univ., Nov. 1972.

______ **and J. C. Panzar**, "Models of Regulated Monopoly With Service Quality and Averch-Johnson Effects: Pre-Empirical Comparative Statics," memo. no. 176, Center Res. Econ. Growth, Stanford, Univ., July 1974.

R. Schmalensee, "Comparative Static Properties of Regulated Airline Oligopolies," *Bell J. Econ.*, Autumn 1977, *8*, 565–76.

L. White, "Quality Variation When Prices Are Regulated," *Bell J. Econ.*, Autumn 1972, *3*, 425–36.

J. V. Yance, "Nonprice Competition in Jet Aircraft Capacity," *J. Indust. Econ.*, Nov. 1972, *21*, 55–71.

Bankruptcy, Limited Liability, and the Modigliani-Miller Theorem

By MARTIN F. HELLWIG*

This paper examines the validity of the Modigliani-Miller theorem in the presence of bankruptcy. The theorem asserts that the value of a firm and the set of return patterns that the capital markets offer to private investors are independent of firm debt-equity ratios. The usual proof of the theorem is based on the presumption that, in perfect capital markets, borrowing by firms and borrowing by individuals can be perfect substitutes.

It is unclear whether this presumption is valid when there is a positive probability that either the firm or the individual who borrows to invest in the firm goes bankrupt. On the one hand, Joseph Stiglitz (1969) and Robert Merton have argued that the Modigliani-Miller theorem remains valid even with bankruptcy if agents who borrow to invest in a firm can limit their liability to the amount of collateral they put up. On the other hand, Vernon Smith (1972) has argued that a margin loan which is secured by a pure equity collateral has a different return pattern from a direct loan to the firm and cannot serve as a substitute for the latter (see also David Baron).[1]

Smith's argument raises a number of issues. First, it brings out the important fact that return patterns for limited liability borrowing and lending depend on the composition of the portfolio that serves as collateral. If a firm goes bankrupt, a pure equity collateral is worthless, whereas a collateral that contains some bonds may still earn a positive return because the bonds have a privileged claim to the firm's remaining assets. In general, margin contracts with different collateral compositions will generate different return patterns, both for the borrower and the lender.

Given this multiplicity of return patterns on margin loans, one would not expect that *all* margin loans can be used as substitutes for direct loans to the firm. This intuition is confirmed by Smith's demonstration that a particular margin loan, namely the loan on pure equity collateral, cannot serve this purpose. The question remains open whether this is actually needed for the Modigliani-Miller theorem. It might be enough, if *some appropriate* margin loan could be used as a substitute for lending to the firm.

In pursuing this problem, one finds a gap in the standard proof of the Modigliani-Miller theorem. Usually one applies an arbitrage argument to show that, as a firm changes its debt-equity ratio, investors in shares and bonds adjust their portfolios so as to leave their overall return patterns unchanged. No such arguments are given for other securities related to the firm; in particular, for margin investments and for margin loans that serve to finance those margin investments. Without an analysis of

*Professor of economics, University of Bonn. Research on this paper was supported by NSF grant SOC 75-13437 at Princeton University and by the Deutsche Forschungsgemeinschaft. I have benefited from the advice of Dwight Jaffee and an anonymous referee.

[1] In a similar vein, Stiglitz (1972) asserts: "The value of the firm decreases because there is a divergence in the estimation of the chances of bankruptcy between the lender and the borrower" (p. 467). This analysis rests upon a form of market segmentation that is without economic merit: In his model, there are two groups of agents, optimists and pessimists. The optimists invest all their wealth in the firm's equity, *valued to make the rate of return equal to that on the riskless asset*. The pessimists are indifferent between the risky firm's bonds and the riskless asset. It follows that the optimists, being more optimistic than the pessimists, must prefer the risky bond to the riskless asset, and hence to the equity. Yet, Stiglitz assumes that the optimists invest in the equity and not in the risky bond. Of course, if one group invests *all* its wealth into one asset, that asset need not satisfy a marginal equality condition at all. A correct analysis of his model leads to the results of this paper.

these securities, the proof of the Modigliani-Miller theorem is simply incomplete.

This point is illustrated by Smith's example of a margin loan with a pure equity collateral. The return pattern on this margin loan depends on the likelihood of firm bankruptcy, which in turn depends on the firm's debt-equity ratio. If the firm raises its debt-equity ratio, it increases the probability that it will go bankrupt and the pure equity collateral will become worthless. The question is whether the lender on such a loan can do anything to neutralize this effect.

The present paper studies these issues and comes to the following conclusions:

i: If only securities issued by firms serve as collateral, the validity of the Modigliani-Miller theorem in the presence of bankruptcy depends on whether or not short sales of all securities are permitted.

ii: If securities issued by individuals as well as securities issued by firms serve as collateral, the Modigliani-Miller theorem is generally valid, even if short sales are prohibited.

However the conditions which ensure the validity of the Modigliani-Miller theorem in the presence of bankruptcy are so strong that the set of return patterns that are available to individual investors under these conditions is practically the same as in a complete system of contingent securities markets. This result suggests that these conditions are subject to similar objections as the assumption that securities markets are complete. In particular, it will be argued that the Modigliani-Miller theorem is invalid if one takes account of moral hazard in loan contracts.

The plan of the paper is as follows: Section I develops the framework of the analysis and formulates the central problem. Section II discusses the arbitrage operations required for the Modigliani-Miller analysis, and the main results are contained in Section III.

I. The Basic Problem

I shall use the standard two-period model of the capital market. In the first period, individual firms buy and sell securities whose returns in the second period are uncertain. The market determines security prices and interest rates to equilibrate demand and supply for all securities. In the second period, a state of the world is realized. Payoffs to security holders are made according to the rules defining the different securities.

A. *Firms' Shares and Bonds*

Firms are financed by shares and bonds. Stockholders and bondholders of a given firm will receive the gross return $\tilde{X}$ that the firm will earn in the second period. Bondholders have priority over stockholders.

A bond has a nominal value of \$1 and a contractual gross return r, which is fixed by the market. If the firm has B bonds outstanding, and its second-period return covers the total contractual obligation rB, bondholders receive \$$r$ per bond. Otherwise the firm goes bankrupt. In this case, its return is divided evenly among the bondholders. The return on a dollar invested in the firm's bonds is

$$(1) \qquad s(\tilde{X}) = \min\left(r, \frac{\tilde{X}}{B}\right)$$

If the firm goes bankrupt, its stockholders earn nothing. If the firm does not go bankrupt, its net return $\tilde{X} - rB$ is distributed evenly among the stockholders. With S shares outstanding, let e be the market price per share and $E = eS$ the market value of the equity. Then the return per share is max $[(\tilde{X} - rB)/S, 0]$. The return per dollar invested in the firm's equity is

$$(2) \qquad d(\tilde{X}) = \max\left(\frac{\tilde{X} - rB}{E}, 0\right)$$

B. *Simple Margin Contracts*

Individual agents invest in shares and bonds. In addition, they can borrow and lend on margin. Under this arrangement, an investor combines borrowed funds with his own money to buy securities. The securities purchased serve as collateral for the loan. In the second period, the borrower's liability to the lender is limited to the contractual repayment or the gross return on the collateral whichever is smaller.

The returns to both the lender and the borrower on a margin contract depend on the margin rate, and on the composition of the portfolio that serves as collateral. If one neglects the costs of writing and administering the contract, one should therefore expect that the collateral composition as well as the margin rate are fixed by the terms of the contract. Margin loans and investments with different margin rates or different collaterals must then be treated as different securities.

A margin contract will be called *simple*, if its collateral consists of shares and bonds of one firm. Throughout most of the paper, I shall assume that all margin contracts are simple. Securities issued by individuals do not serve as collateral. Moreover, agents who want to make margin investments in several firms must take out a different loan for each firm, so that a collateral with securities of one firm does not protect the loan that finances investment in another firm. The restriction to simple margin loans serves to focus the analysis on the original Modigliani-Miller idea that individual agents borrow to invest in a firm.

A simple margin contract is characterized by a pair of numbers (a, k). For every dollar supplied by the lender, the borrower puts up k dollars of his own money. A loan of one dollar contributes to a total investment of $1+k$ dollars that serves as collateral for the loan. The parameter a indicates the proportion of bonds in the collateral. The collateral for a one dollar loan consists of $a(1+k)$ bonds and $(1-a)(1+k)$ dollars worth of stock. It bears the gross return $(1+k)[(1-a)d(\tilde{X})+as(\tilde{X})]$.

Let $\bar{r}(a, k)$ be the contractual repayment on an (a, k) loan. The borrower pays this amount when it is less than the return on the collateral. Otherwise, he simply forfeits the collateral. Thus the per dollar return to the lender is

$$(3) \quad \bar{s}(\tilde{X}; a, k) = \min\{\bar{r}(a, k), \\ (1+k)\left[(1-a)d(\tilde{X})+as(\tilde{X})\right]\}$$

The borrower receives nothing if he forfeits the collateral. Otherwise he receives the difference between the return on the collateral and the repayment to the lender. Since he contributes k dollars for every dollar he borrows, his return on an (a, k) margin investment is

$$(4) \quad \bar{d}(\tilde{X}; a, k) = \frac{1}{k}\max \\ \{(1+k)\left[(1-a)d(\tilde{X})+as(\tilde{X})\right]-\bar{r}(a,k), 0\}$$

C. *Capital Market Equilibrium*

The economy contains F firms and A individual agents, who trade in the markets for firm's shares and bonds, and for simple margin loans. Firm behavior is given exogenously in the simplest possible way. Each firm i chooses a debt-equity ratio $z_i \in \mathbb{R}_+$, which it enforces by issuing bonds to buy back shares if $B_i < z_i E_i$, or issuing additional shares to redeem bonds if $B_i > z_i E_i$.

There is no real investment. The gross return $\tilde{X}_i$ of firm i is exogenous to the analysis. Abstracting from other sources of uncertainty, a state of the world in the second period is then defined as a realization $(X_1, X_2, \ldots, X_F) \in \mathbb{R}^F_+$ of the random vector $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_F)$ of firms returns.

Individual agents use their initial wealth[2] to buy a portfolio of shares, bonds, margin loans, and margin investments. A portfolio provides its owner with a *return pattern* specifying his second-period income as a function of the state of the world. Each agent asks for the portfolio that provides him with the most desirable return pattern he can afford. Preferences over return patterns depend on agents' expectations about the random vector $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_F)$ and on their attitudes towards risk. Preferences are monotone: If the second-period income yielded by one portfolio is at least as high as that yielded by another in every state of the world, and strictly higher in some nonnull set of states of the world, then all agents prefer the first portfolio over the second.

[2] Strictly speaking, their initial holdings of shares and bonds.

Firms and individual agents trade at the *market conditions* given by firms' share prices $e_1, e_2,\dots, e_F$, firms' interest rates $r_1, r_2,\dots, r_F$, and the interest schedules $\bar{r}_i(a,k)$, $a\in[0,1]$, $k\in\mathbb{R}_+$, $i=1,2,\dots,F$ for the (a,k) loan markets connected with firm i. The market conditions correspond to an *equilibrium*, if all markets clear and the following conditions are satisfied:

i: For each firm i, the bond issue B_i and stock issue S_i are compatible with the debt-equity ratio z_i:

$$B_i = z_i e_i S_i \equiv z_i E_i \tag{5}$$

ii: For each firm i, individual agents demand B_i bonds and S_i shares either directly or for loan collaterals.

iii: The demand for loans by borrowers in each (a,k) loan market connected with firm i is equal to the supply of such loans by lenders. Since a borrower must put up k dollars for every dollar he borrows, this means that the demand for (a,k) margin investments is just k times the supply of (a,k) margin loans.

D. *The Modigliani-Miller Principle*

The system of capital markets that we are considering has the distinctive feature that all securities transactions involve the markets for firms' shares and bonds either directly or indirectly when shares and bonds are used as collateral. This contrasts with a complete system of contingent securities markets in which agents buy and sell promises to pay if a specified state of the world occurs without necessarily involving the markets for firm's shares and bonds.

If one had a complete system of contingent securities markets, agents could arrange their portfolios so as to obtain any desired pattern of returns across states of the world. This is not generally true in an incomplete system consisting of markets for firm shares and bonds, and simple margin contracts. Thus, an agent who believes that $\tilde{X}_1=1$ with certainty wants to buy a portfolio whose return is positive if and only if the event $\tilde{X}_1=1$ occurs. But in a system with shares, bonds, and margin contracts, this portfolio may not exist. Then the agent must buy a portfolio whose returns are positive both in the event $\tilde{X}_1=1$ and in some event $\tilde{X}_1\neq 1$, which he does not care for, though other agents do because they have different beliefs.

The question arises what return patterns are available from shares, bonds, and margin contracts. Intuitively, one would expect that the set of available return patterns varies with the debt-equity ratios that are chosen by the firms. Presumably an increase in the debt-equity ratio will raise a firm's bond issue, lower the market valuation of its equity, and require an increase in the contractual interest paid to bondholders. All these changes have a direct impact on the return patterns for shares and bonds, for example, by moving the bankruptcy point $\tilde{X}=rB$ to the right. Since the returns on simple margin contracts are tied to the returns on shares and bonds, one expects that the whole set of available return patterns must be affected.

If changes in the debt-equity ratio do affect the set of available return patterns, one would expect that they also affect the market value of a firm defined as

$$V=B+eS=B+E \tag{6}$$

Presumably the market will value a firm more highly if the set of return patterns that the firm makes available to the market is better suited to accomodate agents' diverse beliefs and preferences over return patterns.

According to Modigliani and Miller, this intuitive reasoning is incorrect. Changes in debt-equity ratios are held to have no effect on either the set of available return patterns or the market valuation of firms. More precisely, the following principle is proposed:

Modigliani-Miller Principle: Let an economy consist of F firms, with gross returns $\tilde{X}_i$, $i=1,2,\dots,F$, and A agents with given preferences over return patterns. Let $(z_1^u, z_2^u,\dots, z_F^u)$, $(z_1^w, z_2^w,\dots, z_F^w)$ be two vectors of debt-equity ratios. Given any equilibrium of the economy with debt-equity ratios $(z_1^u,\dots, z_F^u)$, there exists an equilibrium of the economy with debt-equity ratios $(z_1^w,\dots,z_F^w)$, such that in the two equilibria:

i: The market values of all firms are the same, and

ii: The same patterns of returns across states of the world are available through investment in securities with active markets.

In the rest of the paper I shall analyze under what conditions the Modigliani-Miller principle is valid for a market system with shares, bonds, and margin contracts.

II. Modigliani-Miller Arbitrage

The Modigliani-Miller analysis is based on the argument that those securities whose returns depend directly on firm debt-equity ratios are economically irrelevant. They do not enlarge the set of return patterns available to investors, and moreover, capital market equilibrium would be unaffected if investment in such securities were prohibited. The present section examines the validity of this claim for shares, bonds, and simple margin contracts under bankruptcy.

A. M-M *Contracts*

Those margin contracts whose collaterals contain shares and bonds in the proportions in which firms have issued them form a class of securities whose returns are not directly affected by changes in firm debt-equity ratios.

Given a firm with the value $V=B+E$, the debt-equity ratio $z=B/E$, and the gross return $\tilde{X}$, one defines

$$(7)\qquad \hat{a}(z)=\frac{z}{1+z}=\frac{B}{V}$$

One dollar invested in $\hat{a}(z)$ bonds and $(1-\hat{a}(z))=1/(1+z)=E/V$ dollars worth of shares of the firm bears the return

$$\begin{aligned}&\hat{a}(z)s(\tilde{X})+(1-\hat{a}(z))d(\tilde{X})\\&=\frac{B}{V}\min\left[r,\frac{\tilde{X}}{B}\right]+\frac{E}{V}\max\left[\frac{\tilde{X}-rB}{E},0\right]\\&=\frac{1}{V}[\min(rB,\tilde{X})+\max(\tilde{X}-rB,0)]\\&=\frac{\tilde{X}}{V}\end{aligned}$$

Using (3) and (4), the returns on $(\hat{a}(z),k)$ margin loans and investments are given as

$$(8)\qquad \bar{s}(\tilde{X};\hat{a}(z),k)=\min\left[\bar{r}(\hat{a}(z),k),(1+k)\frac{\tilde{X}}{V}\right]$$

$$(9)\quad \bar{d}(\tilde{X};\hat{a}(z),k)=\frac{1}{k}\max\left[(1+k)\frac{\tilde{X}}{V}-\bar{r}(\hat{a}(z),k),0\right]$$

The returns $\bar{s}(\tilde{X};\hat{a}(z),k)$ and $\bar{d}(\tilde{X};\hat{a}(z),k)$ depend on the interest rate $\bar{r}(\bar{a}(z),k)$ and the firm value V, but not on the debt-equity ratio z. Any effects of changes in z on the returns $d(\tilde{X})$ and $s(\tilde{X})$ on shares and bonds are neutralized by an adjustment in the proportions of shares and bonds serving as collateral. The collateral containing a portion $(1+k)/V$ of all securities issued by the firm simply earns $(1+k)/V$ times the firm return X, no matter how many shares and bonds are outstanding.

In the following, the $(\hat{a}(z),k)$ margin contracts will be referred to as M-M *contracts*. The main issue will be whether all return patterns that are available to individual investors through shares, bonds, and simple margin contracts in equilibrium can also be attained by suitable combinations of M-M margin loans and investments.

Of course, return patterns on the various securities depend on market conditions. One must rely on arbitrage between M-M contracts and other securities to bring market conditions to the point where any attainable return pattern can be attained through M-M contracts only.

B. *Arbitrage for Shares and Bonds*

First, consider arbitrage between a firm's shares and bonds on the one hand, and the $(\hat{a}(z),1/z)$-margin contract on the other. In the $(\hat{a}(z),1/z)$ contract, the ratio $1/k$ of borrowed money to own money is just equal to the debt-equity ratio z. For $k=1/z$, one has $(1+k)/V=(1+z)/zV=1/B$ by (7).

Therefore (8) and (9) imply

$$(10)\quad \bar{s}(\tilde{X};\hat{a}(z),1/z)=\min\left[\bar{r}(\hat{a}(z),1/z),\tilde{X}/B\right]$$

$$(11)\quad \bar{d}(\tilde{X};\hat{a}(z),1/z) =\max\left[\frac{\tilde{X}-\bar{r}(\hat{a}(z),1/z)B}{E},0\right]$$

From (1), (2), (10), and (11), one sees that the return patterns for shares and bonds are the same as for $(\hat{a}(z),1/z)$ margin investments and loans, if $\bar{r}(\hat{a}(z),1/z)=r$. This condition, $\bar{r}(\hat{a}(z),1/z)=r$, is *not* an assumption of the Modigliani-Miller analysis. It follows from the more fundamental condition that loans subject to the risk of default —like all securities—are evaluated in terms of their return patterns over states of the world regardless of the identity of the borrower.

Suppose that $\bar{r}(\hat{a}(z),1/z)>r$. Then no lender will buy bonds, which are dominated by $(\hat{a}(z),1/z)$ margin loans, and no borrower desires $(\hat{a}(z),1/z)$ margin investments, which are dominated by shares. Similarly, if $\bar{r}(\hat{a}(z),1/z)<r$, nobody wants to buy shares which are dominated by $(\hat{a}(z),1/z)$ margin investments, or to make $(\hat{a}(z),1/z)$ margin loans which are dominated by bonds. Therefore one has:

LEMMA 1: *Consider a firm with gross return $\tilde{X}$ and debt-equity ratio z in capital market equilibrium. If there is unlevered investment in the firm's shares and bonds, or if the $(\hat{a}(z),1/z)$-loan market is active, one has*

$$\bar{r}(\hat{a}(z),1/z)=r$$

$$\bar{s}(\tilde{X};\hat{a}(z),1/z)=s(\tilde{X})$$

$$\bar{d}(\tilde{X};\hat{a}(z),1/z)=d(\tilde{X})$$

C. *Return Patterns for Simple Margin Contracts*

The analysis of simple margin contracts is rather more complicated. Their return patterns generally reflect the event of firm bankruptcy as well as default by the borrower. In contrast, the return patterns for M-M contracts reflect only default by the borrower, while return patterns for shares and bonds reflect only firm bankruptcy.

For any given contract, the main question is whether default by the borrower is more or less likely than firm bankruptcy. I shall show that this depends on whether $(1+k)a \gtreqless 1$, i.e., whether the bond holding in the collateral is smaller or larger than the margin loan. The argument is based on the following lemma, which is proved in the Appendix.

LEMMA 2: *Consider a firm with gross return $\tilde{X}$ and debt-equity ratio z in capital market equilibrium. For any active (a,k) loan market related to the firm one has $\bar{r}(a,k)\gtreqless r$ as $(1+k)a\lesseqgtr 1$.*

The point of this lemma is that when both the firm and the (a,k) borrower default, the (a,k) lender receives $(1+k)a\tilde{X}/B$, as opposed to a return $\tilde{X}/B$ from a direct investment in bonds. If $(1+k)a<1$, the difference between $\tilde{X}/B$ and $(1+k)a\tilde{X}/B$ must be compensated by a contractual return $\bar{r}(a,k)>r$. If on the other hand, $(1+k)a\tilde{X}/B>\tilde{X}/B$, the lender must accept a contractual return $\bar{r}(a,k)<r$.

Now $(1+k)a<1$ implies $(1+k)ar<\bar{r}(a,k)$ in equilibrium. If the firm goes bankrupt, the collateral on such a contract earns $(1+k)a\,s(\tilde{X})<(1+k)ar<\bar{r}(a,k)$, and default by the (a,k) borrower follows automatically. Alternatively, $(1+k)a>1$ implies $(1+k)ar>\bar{r}(a,k)$ in equilibrium. Then default by the (a,k) borrower implies $(1+k)a\ \ s(\tilde{X})+(1+k)(1-a)d(\tilde{X})<\bar{r}(a,k)<(1+k)ar$, from which the condition $s(\tilde{X})<r$ for firm bankruptcy follows immediately. This completes the argument that, in equilibrium, default on the (a,k) contract is more or less likely than firm bankruptcy as $(1+k)a<1$ or $(1+k)a>1$.

The return patterns for margin loans and investments with $(1+k)a<1$ and $(1+k)a>1$ are listed in Tables 1 and 2. Also listed are the derivatives of the functions $\bar{s}(\cdot;a,k)$ and $\bar{d}(\cdot;a,k)$. Note that $\bar{s}(X;a,k)$ in-

TABLE 1—RETURNS ON (a, k) SECURITIES WITH $(1+k)a<1$

$\tilde{X}$	$\in[0, rB]$	$\in\left[rB, rB+E\frac{\bar{r}(a,k)-(1+k)ar}{(1+k)(1-a)}\right]$	$\in\left[rB+E\frac{\bar{r}(a,k)-(1+k)ar}{(1+k)(1-a)}, \infty\right)$
$\bar{s}(\tilde{X}; a, k)$	$(1+k)a\frac{\tilde{X}}{B}$	$(1+k)[(1-a)\frac{\tilde{X}-rB}{E}+ar]$	$\bar{r}(a,k)$
$\bar{d}(\tilde{X}; a, k)$	0	0	$\frac{1}{k}\{(1+k)\left[(1-a)\frac{\tilde{X}-rB}{E}+ar\right]-\bar{r}(a,k)\}$
$d\bar{s}/d\tilde{X}$	$(1+k)\frac{a}{b}$	$(1+k)\frac{1-a}{E}$	0
$d\bar{d}/d\tilde{X}$	0	0	$\frac{1+k}{k}\frac{1-a}{E}$

TABLE 2—RETURNS ON (a, k) SECURITIES WITH $(1+k)a>1$

$\tilde{X}$	$\in\left[0, \frac{\bar{r}(a,k)}{(1+k)a}B\right]$	$\in\left[\frac{\bar{r}(a,k)}{(1+k)a}B, rB\right]$	$\in[rB. \infty)$
$\bar{s}(\tilde{X}; a, k)$	$(1+k)a\frac{\tilde{X}}{B}$	$\bar{r}(a,k)$	$\bar{r}(a,k)$
$\bar{d}(\tilde{X}; a, k)$	0	$\frac{1}{k}\left[(1+k)a\frac{\tilde{X}}{B}-\bar{r}(a,k)\right]$	$\frac{1}{k}\left\{(1+k)\left[(1-a)\frac{\tilde{X}-rB}{E}+ar\right]-\bar{r}(a,k)\right\}$
$d\bar{s}/d\tilde{X}$	$(1+k)\frac{a}{b}$	0	0
$d\bar{d}/d\tilde{X}$	0	$\frac{1+k}{k}\frac{a}{b}$	$\frac{1+k}{k}\frac{1-a}{E}$

creases in X up to the point of borrower default and then remains constant at $\bar{r}(a, k)$. In contrast, $\bar{d}(X; a, k)$ remains constant at 0 up to the point of borrower default and then increases with X. Thus all return patterns exhibit kinks at the point of borrower default at which $(1+k)a \; s(\tilde{X})+(1+k)(1-a)\bar{d}(\tilde{X})=\bar{r}(a, k)$.

In addition, the functions $\bar{s}(\cdot; a, k)$ with $(1+k)a<1$ have a kink at the point $\tilde{X}=rB$ of firm default when their slopes switch from $(1+k)a/B$ to $(1+k)(1-a)/E$. Similarly, at $\tilde{X}=rB$, the slopes of the functions $\bar{d}(\cdot; a, k)$ with $(1+k)a>1$ switch from $(1+k)a/kB$ to $(1+k)(1-a)/kE$. This kink at the point $\tilde{X}=rB$ does not appear in M-M contracts because $\hat{a}(z)/B=(1-\hat{a}(z))/E$. For all but M-M contracts, either $\bar{s}(\cdot; a, k)$ or $\bar{d}(\cdot; a, k)$ has a kink at $\tilde{X}=rB$. Thus the return patterns in all but M-M contracts are directly affected by changes in the firm's financial policy.

D. *Arbitrage for Simple Margin Contracts*

Let us consider the possibility of arbitrage between M-M contracts and other margin contracts. Return patterns for M-M contracts have a simpler structure than return patterns for other margin contracts. Whereas return patterns for other margin loans and investments have two kinks at the point of borrower default and of firm bankruptcy, return patterns for M-M contracts have only a single kink at the point of borrower default. Therefore one needs two M-M contracts to replicate one other margin contract.

The basic approach is illustrated in Figure 1. This figure shows the return patterns for an (a, k) loan, $\bar{s}(\tilde{X}; a, k)$, with $(1+k)a<1$ and $a>\hat{a}(z)$, as well as for the M-M loan whose return pattern has a kink at the point of firm bankruptcy and the M-M loan

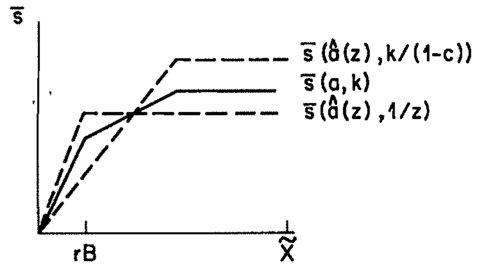


FIGURE 1

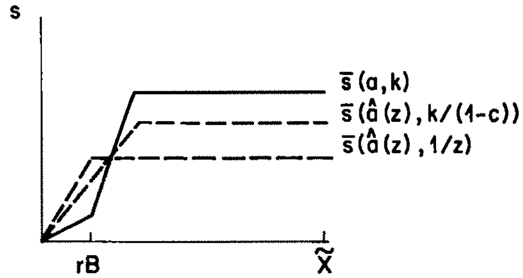


FIGURE 2

whose return pattern has a kink at the point of (a, k) borrower default.

Figure 1 is drawn so that all three return patterns intersect in the same point. In this case the return pattern on the (a, k) loan is a convex combination of the return patterns on the two M-M loans. This means that there exists some combination of the two M-M loans whose return pattern is $\bar{s}(\tilde{X}; a, k)$.

The key fact that all three return patterns intersect in the same point must be established by arbitrage. Roughly, $\bar{s}(\tilde{X}; a, k)$ is dominated by a combination of the two M-M loans if $\bar{r}(a, k)$ is so low that $\bar{s}(\tilde{X}; a, k)$ passes below the intersection point of the other two contracts. If, on the other hand, $\bar{s}(\tilde{X}; a, k)$ passes above that intersection point, then $\bar{r}(a, k)$ is so high that $\bar{d}(\tilde{X}; a, k)$ is dominated by some M-M margin investment. In neither case can the (a, k) loan market be active.

A difficulty arises when $(1+k)a<1$ and $a<\hat{a}(z)$. In this case, $a/B<(1-a)/E$, and the return pattern $\bar{s}(\tilde{X}; a, k)$ has the shape shown in Figure 2. Now the function $\bar{s}(\cdot\,; a, k)$ is not concave. Therefore it cannot be obtained as a convex combination of return patterns on M-M loans, which are all concave functions. Any replication of $\bar{s}(\tilde{X}; a, k)$ by M-M loans must involve negative holdings, that is, short sales of one M-M loan. Then the feasibility of Modigliani-Miller arbitrage depends on whether or not short sales are permitted.

Formally, one has:

LEMMA 3: *Suppose that short sales of all securities are permitted and consider a firm with gross return $\tilde{X}$ and debt-equity ratio z in capital market equilibrium. For any active (a, k) margin contract, one has*

a: If $(1+k)a<1$, then

$$\bar{r}(a, k)=c\bar{r}\left(\hat{a}(z), \frac{1}{z}\right)+(1-c)\bar{r}\left(\hat{a}(z), \frac{k}{1-c}\right)$$

$$\bar{s}(\tilde{X}; a, k)=c\bar{s}\left(\tilde{X}; \hat{a}(z), \frac{1}{z}\right)+(1-c)\bar{s}\left(\tilde{X}; \hat{a}(z), \frac{k}{1-c}\right)$$

$$\bar{d}(\tilde{X}; a, k)=\bar{d}\left(\tilde{X}; \hat{a}(z), \frac{k}{1-c}\right)$$

where $c=(1+k)(1+z)(a-\hat{a}(z))$

b: If $(1+k)a>1$, then

$$\bar{r}(a, k)=\bar{r}(\hat{a}(z),(1-\gamma)k)$$

$$\bar{s}(\tilde{X}; a, k)=\bar{s}(\tilde{X}; \hat{a}(z),(1-\gamma)k)$$

$$\bar{d}(\tilde{X}; a, k)=\gamma\bar{d}\left(\tilde{X}; \hat{a}(z), \frac{1}{z}\right)+(1-\gamma)\bar{d}(\tilde{X}; \hat{a}(z),(1-\gamma)k)$$

where $\gamma=\dfrac{1+k}{k}\left[1-\dfrac{a}{\hat{a}(z)}\right]$

PROOF: (See the Appendix.)

LEMMA 4: *Margin contracts with $(1+k)a<1$ and $a<\hat{a}(z)$, or $(1+k)a>1$ and $a>\hat{a}(z)$,* *cannot be replicated by M-M contracts if short sales are prohibited.*

PROOF: (See the Appendix.)

In summary, Modigliani-Miller arbitrage is generally feasible, if short sales of all securities are permitted. Under this condition any portfolio whose return pattern depends explicitly on the firm's debt policy can be duplicated by a portfolio of M-M loans and investments whose return pattern is not directly affected by changes in the debt-equity ratio z.

But Modigliani-Miller arbitrage is infeasible for some margin contracts if short sales are not allowed. If there is active borrowing and lending in such contracts, one would expect the Modigliani-Miller principle to break down because changes in the debt-equity ratio affect the return patterns for such contracts.

III. Unrestricted Margin Contracts and the Modigliani-Miller Theorem

A. *The Irrelevance of Shares, Bonds, and Non-*M-M *Contracts*

Securities subject to Modigliani-Miller arbitrage are irrelevant for market equilibrium as well as for the set of available return patterns. If one prohibits investment in such securities, agents will simply shift to the corresponding M-M contracts. This shift has no effect on market equilibrium. All markets that were in equilibrium before are still in equilibrium after the shift to M-M contracts.

LEMMA 5: *Consider an economy with shares, bonds, and simple margin contracts in capital market equilibrium, and let short sales of all securities be permitted. Market equilibrium is preserved and the allocation of return patterns is unchanged, if all individual agents are restricted to invest in* M-M *loans or margin investments.*

PROOF:

I shall proceed in two steps: In the first, agents are restricted to shares, bonds, and M-M contracts; in the second, to M-M contracts only.

Step 1: Consider the (a,k) margin contract connected with firm i, where $(1+k)a < 1$. Suppose that, in the initial equilibrium, y \$ are placed in (a,k) loans and $x=ky$ \$ in (a,k) margin investments. Lemmas 1 and 3 imply

$$\bar{s}_i(\tilde{X}_i;a,k)=cs_i(\tilde{X}_i)$$
$$+(1-c)\bar{s}_i\left(\tilde{X}_i;\hat{a}(z_i),\frac{k}{1-c}\right)$$
$$\bar{d}_i(\tilde{X}_i;a,k)=\bar{d}_i\left(\tilde{X}_i;\hat{a}(z_i),\frac{k}{1-c}\right)$$

where $\quad c=(1+k)(1+z_i)(a-\hat{a}(z_i))$

If the (a,k) margin loan is abolished, the initial y \$ in (a,k) loans are shifted to cy \$ in bonds and $(1-c)y$ \$ in $(\hat{a}(z_i),k/(1-c))$ loans. Similarly, the initial x \$ in (a,k) margin investments are shifted to x \$ in $(\hat{a}(z_i),k/(1-c))$ margin investments generating a demand for $(1-c)x/k$ \$ in additional margin loans. Since $x=ky$, this additional demand is exactly matched by the additional supply of $(1-c)y$ \$ in $(\hat{a}(z_i),k/(1-c))$ margin loans.

This shift leaves total share and bond holdings unaffected. Instead of $(1-a)(1+k)y$ \$ worth of equity in (a,k) loan collaterals, agents now hold $(1-\hat{a}(z_i))(1+k/(1-c))(1-c)y$ \$ worth of equity in $(\hat{a}(z_i),k/(1-c))$ loan collaterals. Instead of $a(1+k)y$ bonds in (a,k) loan collaterals, they hold cy bonds directly and $\hat{a}(z_i)(1+k/(1-c))(1-c)y$ bonds in $(\hat{a}(z_i),k/(1-c))$ loan collaterals. It is easy to verify that total share and bond holdings before and after the abolition of the (a,k) contract are the same.

Essentially the same argument can be applied to the case $(1+k)a>1$. This is left to the reader.

Step 2: Suppose that, in addition to investment in non-M-M contracts, direct investment in shares and bonds is also prohibited. From Lemma 1, $s_i(\tilde{X}_i)=\bar{s}_i(\tilde{X}_i;\hat{a}(z_i),1/z_i)$, and $d_i(\tilde{X}_i)=\bar{d}_i(\tilde{X}_i;\hat{a}(z_i),1/z_i)$, so every dollar that was initially in bonds is now put into $(\hat{a}(z_i),1/z_i)$

margin loans, and every dollar that was initially in shares is now put into $(\hat{a}(z_i), 1/z_i)$ margin investments. As in Step 1, one can show that this shift has no effect on equilibrium in the markets for shares, bonds, and $(\hat{a}(z_i), 1/z_i)$ loan contracts. The argument rests on the fact that, in the absence of non-M-M contracts, the ratio of unlevered holdings of bonds to unlevered holdings of shares is z_i because $B_i = z_i E_i$, and, moreover, all collaterals hold bonds and shares in the proportion $(\hat{a}(z_i)/(1-\hat{a}(z_i)) = z_i$. The details are left to the reader.

B. *Short Sales and the Modigliani-Miller Theorem*

The irrelevance of firm debt-equity ratios follows directly from Lemma 5. If all shares and bonds are held in M-M loan collaterals, than any change in debt-equity ratios is neutralized by the corresponding change in collateral compositions. Therefore one has

PROPOSITION 1: *An economy with shares, bonds, and simple margin contracts in which short sales of all securities are permitted satisfies the Modigliani-Miller principle.*

PROOF:

Let $(z_1^u, \dots, z_F^u)$, $(z_1^w, \dots, z_F^w)$ be two vectors of debt-equity ratios for firms $1, 2, \dots, F$. Further, let E_i^u, r_i^u, $\bar{r}_i(a,k)$, $i = 1, 2, \dots, F$, $a \in [0,1]$, $k \geqslant 0$ be a set of equilibrium market conditions for the debt-equity ratios $(z_1^u, \dots, z_F^u)$. Define new market conditions corresponding to $(z_1^w, \dots, z_F^w)$ by the equations

$$E_i^w = \frac{1+z_i^u}{1+z_i^w} E_i^u \qquad i = 1, 2, \dots, F \tag{12}$$

$$\bar{r}_i^w(\hat{a}(z_i^w), k) = \bar{r}_i^u(\hat{a}(z_i^u), k) \qquad i = 1, 2, \dots, F,\ k \geqslant 0 \tag{13}$$

together with the arbitrage relations of Lemmas 1 and 3.

Then (12) implies $V_i^w = B_i^w + E_i^w = (1+z_i^w)E_i^w = (1+z_i^u)E_i^u = V_i^u$. Further, using (12) and (13) in (8) and (9), one has

$$\bar{s}_i^w(\tilde{X}_i; \hat{a}(z_i^w), k) = \bar{s}_i^u(\tilde{X}_i; \hat{a}(z_i^u), k)$$

and $\quad \bar{d}_i^w(\tilde{X}_i; \hat{a}(z_i^w), k) = \bar{d}_i^u(\tilde{X}_i; \hat{a}(z_i^u), k)$

for all i, k. Thus in both situations the same return patterns are available through M-M contracts. From Lemmas 1 and 3, the set of available return patterns is unaffected by the change in debt-equity ratios. Therefore, agents' preferred return patterns in the two situations are identical. Instead of y \$ in $(\hat{a}(z_i^u), k)$ margin loans, and x \$ in $(\hat{a}(z_i^u), k)$ margin investments, they now place y \$ in $(\hat{a}(z_i^w), k)$ margin loans and x \$ in $(\hat{a}(z_i^w), k)$ margin investments. It is then easy to verify that the market conditions (12) and (13) correspond to an equilibrium for the debt-equity ratios $(z_1^w, \dots, z_F^w)$, if and only if one originally had an equilibrium for the debt-equity ratios $(z_1^u, \dots, z_F^u)$.

The assumption that short sales are permitted is necessary as well as sufficient for the validity of the Modigliani-Miller principle. As we have seen, without short sales, the arbitrage operation of Lemma 3 breaks down. Therefore one has

PROPOSITION 2: *For an economy with shares, bonds, and simple margin contracts in which no short sales are permitted, the following statements are equivalent:*

i: *The Modigliani-Miller principle is valid;*

ii: *For any vector of debt-equity ratios $(z_1, \dots, z_F)$ only margin contracts with $(1+k)a < 1$ and $a \geqslant \hat{a}(z_i)$ or $(1+k)a > 1$ and $a \leqslant \hat{a}(z_i)$ are active in connection with firm i.*

PROOF:

For margin contracts with $(1+k)a < 1$ and $a \geqslant \hat{a}(z_i)$, or $(1+k)a > 1$ and $a \leqslant \hat{a}(z_i)$, the arbitrage argument of Lemma 3 does not require short sales. In this case the argument leading up to Proposition 1 remains valid because short sales are not used anywhere else.

To prove implication (i)⇒(ii), note first that the Modigliani-Miller principle implies: $r_i^u B_i^u \gtreqless r_i^w B_i^w$ as $z_i^u \gtreqless z_i^w$, i.e. an increase in the debt-equity ratio moves the point of firm bankruptcy to the right. If one had $z_i^u < z_i^w$ and $r_i^u B_i^u \geqslant r_i^w B_i^w$, then one would have $B_i^u < B_i^w$ and $r_i^u > r_i^w$ and $s_i^u(\tilde{X}_i) >$

$s_i^w(\tilde{X}_i)$ for all $\tilde{X}_i$. The Modigliani-Miller replication of $s_i^u(\tilde{X}_i)$ would dominate $s_i^w(\tilde{X}_i)$, and one could not have an equilibrium.

Now suppose that the M-M principle is valid and that in some equilibrium with debt-equity ratios $(z_1^u,\dots,z_F^u)$, an (a,k) loan market with $(1+k)a<1$ and $a<\hat{a}(z_i)$ is active. The return pattern $\bar{s}_i^u(\tilde{X}_i; a,k)$ is convex at the point of firm bankruptcy $\tilde{X}_i = r_i^u B_i^u$. If firm i's debt-equity ratio rises to $z_i^w > z_i^u$, the return pattern $\bar{s}_i^u(\tilde{X}_i; a,k)$ must be replicated by a portfolio of margin loans whose return pattern is also convex at $\tilde{X}_i = r_i^u B_i^u$. Since short sales are prohibited and return patterns on margin loans are now convex *only* at the point $\tilde{X}_i = r_i^w B_i^w$, one must have $r_i^w B_i^w = r_i^u B_i^u$, which is impossible. A similar argument applies to the return pattern $\bar{d}_i^u(\tilde{X}_i; a,k)$ with $(1+k)a>1$ and $a>\hat{a}(z_i^u)$. This completes the proof of Proposition 2.

It is of some interest to note that the bounds in condition (ii) are always violated if $a=0$ or $a=1$. Therefore Proposition 2 has the immediate[3]

COROLLARY: *In an economy with shares, bonds, and simple margin contracts in which no short sales are permitted, the Modigliani-Miller principle is invalid, if there is active borrowing and lending on pure equity or pure bond collateral.*

Condition (ii) in Proposition 2 should be regarded as a statement about the diversity of opinion in the market. If opinions are very diverse, one would expect this condition to fail. Thus, two agents who are both more optimistic about a firm than the rest of the market are likely to agree that the collateral should contain more equity than the market portfolio. The margin contract itself then allows them to bet on their remaining differences of opinion.

For example consider a firm that has issued 1 million bonds at $r=1.1$, if the equality is valued at $1 million, two agents with point expectations of $\tilde{X}=2.5$ million and $\tilde{X}=3$ million will wish to conclude a $(0,k)$ loan contract with $\bar{r}(a,k)\in(1.4,1.9)$.

In summary, short sales are necessary as well as sufficient, if the Modigliani-Miller principle is to hold for all possible constellations of beliefs and preferences.

C. *Unrestricted Margin Contracts and the Modigliani-Miller Theorem*

The breakdown of the Modigliani-Miller theorem in the absence of short sales is due to the assumption that securities issued by individuals do not serve as collateral. To see this, consider the pure equity contract, which cannot be replicated by M-M contracts unless short sales are permitted. The collateral for the $(0,k)$ margin contract related to firm i has the return pattern $(1+k)d_i(\tilde{X}_i)$. By Lemma 1, $(1+k)$ dollars placed in $(\hat{a}(z_i), 1/z_i)$ margin investments would have exactly the same return pattern: $(1+k)\bar{d}_i(\tilde{X}_i; \hat{a}(z_i), 1/z_i) = (1+k)d_i(\tilde{X}_i)$. Therefore a margin loan secured by $(1+k)$ dollars of $(\hat{a}(z_i), 1/z_i)$ margin investments should have the same return pattern as the $(0,k)$ loan. Moreover, the return pattern on such a margin loan can be made independent of the firm's debt-equity ratio, because any change in the debt-equity ratio is neutralized by an appropriate change in the composition of the collateral for the margin investment that in turn serves as collateral for the margin loan in question. Therefore, the absence of margin contracts that use $(\hat{a}(z_i), 1/z_i)$ margin investments as collateral is crucial in Proposition 2.

In the absence of transactions costs and moral hazard, one would presume that both the lender and the borrower on a margin contract worry only about the return pattern on the collateral. The composition of the collateral should matter only if it affects the return pattern. From this point of view, the restriction to simple margin contracts would seem to be unwarranted. Moreover, a borrower and a lender should be able to take *any* security that is available in the market

[3] One can use the argument behind Lemma 3 to show that every simple margin contract can be replicated by a combination of $(0,k)$, $(1,k)$, and $(\hat{a}(z),k)$ loan markets. Therefore, borrowing and lending on pure equity or pure bond collateral is in some sense necessary as well as sufficient for the breakdown of Modigliani-Miller arbitrage.

and use it as collateral. Formally, the set of available margin contracts will be called *unrestricted*, if it satisfies condition (14):

(14) Given $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, let the return pattern $f(\tilde{X}_i)$ be available in the market at an expense of one dollar. Then, for every $k>0$, there exists an (f, k) margin contract, with the contractual interest rate $\bar{\bar{r}}_i(f, k)$ and the per dollar return patterns:

$$\bar{\bar{s}}_i(\tilde{X}_i; f, k) = \min\left[(1+k)f(\tilde{X}_i), \bar{\bar{r}}_i(f, k)\right]$$

$$\bar{\bar{d}}_i(\tilde{X}_i; f, k) = \frac{1}{k}\max\left[(1+k)f(\tilde{X}_i) - \bar{\bar{r}}_i(f, k), 0\right]$$

If the capital markets satisfy condition (14), the market conditions of the economy are given by the values of firms' equities, $E_1, E_2, \ldots, E_F$, firms' interest rates $r_1, r_2, \ldots, r_F$, and interest schedules $\bar{\bar{r}}_i(f, k)$ for the (f, k) loan markets connected with firms $i = 1, 2, \ldots, F$.

Under condition (14), the interest rate and the return patterns for a margin contract depend on the composition of the collateral only as it affects the return pattern of the collateral. Therefore, borrowers and lenders will be indifferent, if all shapes and bonds in collaterals are replaced by $(\hat{a}(z_i), 1/z_i)$ margin investments and loans. By the same reasoning that was used in Lemma 5, this move would not upset capital market equilibrium. With unrestricted margin contracts, the market can always buy the firm wholesale and then use the relevant set of available return patterns. Formally, one has:

LEMMA 6: *Consider an economy with markets for shares, bonds, and unrestricted margin contracts in capital market equilibrium. Market equilibrium is preserved and the allocation of return patterns is unchanged, if all portfolios are constrained to hold firm's shares and bonds in the proportions in which they are issued.*

The irrelevance of corporate financial policy follows immediately:

PROPOSITION 3: *The Modigliani-Miller principle is valid in an economy with markets for shares, bonds, and unrestricted margin contracts.*

The proof of Lemma 6 and Proposition 3 is left to the reader.

D. *The Set of Available Return Patterns with Short Sales or Unrestricted Margin Contracts*

The conditions ensuring the validity of the Modigliani-Miller principle are very strong. These conditions may be appropriate for a system of "perfect" capital markets in which there are no frictions. But such a system may have no more practical relevance than a complete system of contingent securities markets.

In fact, if short sales are allowed, the set of available return patterns is approximately the same as in a complete system of Arrow-Debreu securities on the "marginal events" $\tilde{X}_i \in A \subset \mathbb{R}_+$. Adapting an argument of Sanford Grossman and Stiglitz, one obtains

PROPOSITION 4: *Consider an economy with shares, bonds, and simple margin contracts, and let short sales of all securities be permitted. For any measurable function $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $f(0) = 0$, the return pattern $f(\tilde{X}_i)$ lies in the closure of the set of return patterns that are available at the equilibrium market conditions.*

PROOF: (See the Appendix.)

The set of available return patterns is somewhat less rich if short sales are ruled out, but the system of margin contracts is unrestricted. In this case, the set of available return patterns includes approximately every *nondecreasing* function of the firm return X_i. The difference is due to the fact that $s_i(X_i)$ and $d_i(X_i)$ and therefore all return patterns generated by margin borrowing and lending are nondecreasing. Formally one has:

PROPOSITION 5: *Consider an economy with shares, bonds, and unrestricted margin* *contracts. For any nondecreasing function $f:\mathbb{R}_+\to\mathbb{R}_+$ with $f(0)=0$, the return pattern $f(\tilde{X}_i)$ lies in the closure of the set of return patterns that are available at the equilibrium market conditions.*

PROOF: (See the Appendix.)

Propositions 4 and 5 contradict the intuitive notion that a system of markets for shares, bonds, and margin contracts is much simpler than a system of complete contingent securities markets. They suggest that unrestricted short sales and margin contracts are beset with the same difficulties that account for the incompleteness of markets in the first place.

In particular, one must consider the moral hazard that agents fail to fulfill their obligations from credit contracts. The institution of lending on collateral reduces this problem somewhat. But it raises the new difficulty that whoever administrates the collateral might prefer to embezzle it rather than share the returns with his partners. To deal with this difficulty, agents will have to incur transactions and monitoring costs.

These considerations suggest that Propositions 1 and 3 cannot be taken to ensure the validity of the Modigliani-Miller principle in practice. Like the Modigliani-Miller theorem for Arrow-Debreu markets (Stiglitz, 1969; Baron) these propositions deal with an abstract system which ignores essential features of actual capital markets. From a practical point of view, it seems reasonable to suppose that the Modigliani-Miller principle fails when there is a chance of bankruptcy.

Appendix

PROOF of Lemma 2:

First let $(1+k)a<1$ and suppose that $\bar{r}(a,k)\leqslant r$, contrary to the lemma. Then the (a,k) margin loan is dominated by a direct investment in bonds: For $\tilde{X}\leqslant rB$, one has $d(\tilde{X})=0$, and therefore $\bar{s}(\tilde{X};\, a,\, k)=\min[(1+k)as(\tilde{X}),\bar{r}(a,k)]\leqslant \min[(1+k)as(\tilde{X}),r]=(1+k)\,as\,(\tilde{X})<\ s(\tilde{X})$; for $\tilde{X}>rB$, one has $\bar{s}(\tilde{X};a,k)\leqslant\bar{r}(a,k)\leqslant r$. Therefore no (a,k) loans are made, contrary to the assumption that the (a,k) loan market is active. Hence, $(1+k)a<1$ implies $\bar{r}(a,k)>r$.

Similarly, if $(1+k)a>1$ and $\bar{r}(a,k)\geqslant r$, the (a,k) margin investment is dominated by an unlevered holding of $(1+k)(1-a)/k$ dollars in shares and $[(1+k)a-1]/k$ dollars in bonds. One has

$$\bar{d}(\tilde{X};a,k)\leqslant\frac{1}{k}\big[(1+k)(1-a)d(\tilde{X})$$
$$+((1+k)a-1)s(\tilde{X})\big]$$

with strict inequality if $\tilde{X}<rB$. Again the (a,k) loan market cannot be active. Activity of the (a,k) loan market and $(1+k)a>1$ imply $\bar{r}(a,k)<r$.

PROOF of Lemma 3:

With short sales, the argument of Lemma 1 shows that $\bar{r}(\hat{a}(z),1/z)=r$. Now if $(1+k)a<1$, then $\bar{r}(a,k)=c\bar{r}(\hat{a}(z),1/z)+(1-c)\bar{r}(\hat{a}(z),k/(1-c))$ implies $\bar{s}(\tilde{X};a,k)=c\bar{s}(\tilde{X};\hat{a}(z),1/z)+(1-c)\bar{s}(\tilde{X};\hat{a}(z),k/(1-c))$, and $d(\tilde{X};a,k)=\bar{d}(\tilde{X};\hat{a}(z),k/(1-c))$, as the reader may verify by a tedious, but elementary, calculation. By the same calculation, $\bar{r}(a,k)<c\bar{r}(\hat{a}(z),1/z)+(1-c)\bar{r}(\hat{a}(z),k/(1-c))$ implies $\bar{s}(\tilde{X};a,k)\geqslant c\bar{s}(\tilde{X};\hat{a}(z),1/z)+(1-c)\bar{s}(\tilde{X};\hat{a}(z),k/(1-c))$ with strict inequality for large realizations of $\tilde{X}$, and there is no investment in (a,k) loans. Conversely, $\bar{r}(a,k)>c\bar{r}(\hat{a}(z),1/z)+(1-c)\bar{r}(\hat{a}(z),k/(1-c))$ implies $\bar{d}(\tilde{X};a,k)\leqslant\bar{d}(\tilde{X};\hat{a}(z),k/(1-c))$, with strict inequality for large realizations of $\tilde{X}$, so there is no demand for (a,k) margin investments. This proves part *a* of the lemma. Part *b* follows by essentially the same argument.

PROOF of Lemma 4:

First let $(1+k)a<1$ and consider the return pattern $\bar{s}(\tilde{X};a,k)$. Clearly, a portfolio that replicates $\bar{s}(\tilde{X};a,k)$ cannot contain margin investments.

Now the return pattern on any M-M loan is concave in $\tilde{X}$, by (8). Therefore the return pattern on any portfolio of M-M loans that does not involve short sales is concave in $\tilde{X}$. If the return pattern $\bar{s}(\tilde{X};a,k)$ can be replicated by a portfolio of M-M loans, it must then also be concave in $\tilde{X}$. By inspection of Table 1, this implies $a\ntriangleleft\hat{a}(z)$.

Similarly, in the absence of short sales the return pattern on any portfolio of M-M margin investments is convex in $\tilde{X}$, by (9). By inspection of Table 2, it follows that such a portfolio cannot replicate a return pattern $\tilde{d}(\tilde{X}; a, k)$ with $(1+k)a>1$ and $a>\hat{a}(z)$.

PROOF of Proposition 4:

Given a set $A \subset \mathbb{R}_+$, let χ_A be the indicator function of A. Since a measurable function can be approximated by a sequence of simple functions, it suffices to prove the proposition for the return patterns $\chi_A(\tilde{X}_i)$. In fact, because any measurable set A is a countable union of differences of intervals, it suffices to prove the proposition for the return patterns $\chi_{[a,b]}(\tilde{X}_i)$, where $a>0$, $b \geqslant a$. This is done by the following construction.

First, note that the equilibrium interest rate $\bar{r}_i(\hat{a}(z_i), k)$ on M-M loans decreases with k.[4] Therefore one can associate with each $X \in \mathbb{R}_+$ exactly one value $k_i(X)$, so that X is the default point of the $(\hat{a}(z_i), k_i(X))$ loan, i.e., $(1+k_i(X))X=\bar{r}_i(\hat{a}(z_i), k_i(X))V_i$.

For any $0<X^1<X^2<X^3<X^4$, define $k_i^1>k_i^2>k_i^3>k_i^4$ and $r_i^1<r_i^2<r_i^3<r_i^4$ by the relations $k_i^j=k_i(X^j)$ and $r_i^j=\bar{r}_i(\hat{a}(z_i), k_i(X^j))$. Now consider a portfolio of $-X^1/r_i^1(X^2-X^1)$ dollars in $(\hat{a}(z_i), k_i^1)$ loans, $X^2/r_i^2(X^2-X^1)$ dollars in $(\hat{a}(z_i), k_i^2)$ loans, $X^3/r_i^3(X^4-X^3)$ dollars in $(\hat{a}(z_i), k_i^3)$ loans, and $-X^4/r_i^4(X^4-X^3)$ dollars in $(\hat{a}(z_i), k_i^4)$ loans. This portfolio has the return pattern:

$$h(\tilde{X}_i)=\begin{cases} 0 & \text{if} \quad \tilde{X}_i \in [0, X^1] \\ \dfrac{\tilde{X}_i-X^1}{X^2-X^1} & \text{if} \quad \tilde{X}_i \in [X^1, X^2] \\ 1 & \text{if} \quad \tilde{X}_i \in [X^2, X^3] \\ \dfrac{X^4-\tilde{X}_i}{X^4-X^3} & \text{if} \quad \tilde{X}_i \in [X^3, X^4] \\ 0 & \text{if} \quad \tilde{X}_i \in [X^4, \infty) \end{cases}$$

If $X^2=a$, $X^3=b$, $X^1 \to X^2$, $X^4 \to X^3$, then $h(\tilde{X}_i) \to \chi_{[a,b]}(\tilde{X}_i)$, pointwise.

Remark: The return pattern h in the preceding proof has the cost:

$$\left[\frac{X^2}{r_i^2}-\frac{X^1}{r_i^1}\right]\frac{1}{X^2-X^1}$$
$$-\left[\frac{X^4}{r_i^4}-\frac{X^3}{r_i^3}\right]\frac{1}{X^4-X^3}$$

As $X^1 \to X^2$ and $X^4 \to X^3$, h decreases monotonically. Therefore, its cost decreases monotonically to the limit:

$$\text{(A1)} \qquad \lim_{X \nearrow X^2} \frac{d}{dX}\frac{X}{\bar{r}_i(\hat{a}(z_i), k_i(X))} - \lim_{X \searrow X^3} \frac{d}{dX}\frac{X}{\bar{r}_i(\hat{a}(z_i), k_i(X))}$$

This expression (A1) represents the implicit Arrow-Debreu price of the return pattern $\chi_{[a,b]}(\tilde{X}_i)$ with $X^2=a$, $X^3=b$.

PROOF of Proposition 5:

The argument proceeds in two steps:

Step 1: Let $f: \mathbb{R}_+ \to \mathbb{R}_+$ be piecewise linear, continuous, and nondecreasing, with $f(0)=0$. Then the return pattern $f(\tilde{X}_i)$ is available to investors in an economy with shares, bonds, and unrestricted margin contracts.

Let $X^0=0, X^1, X^2, \ldots$ be the sequence of kinks of f. One has $f=\Sigma_{j=0}^{\infty} g^j$, where the functions $g^0, g^1, g^2, \ldots$ are given as

$$g^j(X)=\begin{cases} 0 & \text{if } X \in [0, X^j] \\ (X-X^j)\dfrac{f(X^{j+1})-f(X^j)}{X^{j+1}-X^j} & \\ & \text{if } X \in [X^j, X^{j+1}] \\ f(X^{j+1})-f(X^j) & \text{if } X \geqslant X^{j+1} \end{cases}$$

The following facts are easily verified:

a: If $X^j=0$, $X^{j+1}=\infty$, the return pattern

[4] $k_1>k_2$ and $\bar{r}_i(\hat{a}(z_i), k_1)>\bar{r}_i(\hat{a}(z_i), k_2)$ would imply $\bar{s}_1(X_i; \hat{a}(z_i), k_1)>\bar{s}_i(X_i; \hat{a}(z_i), k_2)$, with strict inequality, if $(1+k_1)X_i<\bar{r}_i(\hat{a}(z_i), k_1)V_i$.

$g^j(\tilde{X}_i)$ is proportional to $\tilde{X}_i/V_i = \hat{a}_i s_i(\tilde{X}_i)+(1-\hat{a}_i)d_i(\tilde{X}_i)$.

b: If $X^j=0$, $X^{j+1}<\infty$, the return pattern $g^j(\tilde{X}_i)$ is proportional to $\bar{\bar{s}}_i(\tilde{X}_i; \hat{a}_i s_i + (1-\hat{a}_i)d_i, k_i^j)$ for some k_i^j.

c: If $X^j>0$, $X^{j+1}=\infty$, the return pattern $g^j(\tilde{X}_i)$ is proportional to $\bar{\bar{d}}(\tilde{X}_i; \hat{a}_i s_i + (1-\hat{a}_i)d_i, k_i^j)$ for some k_i^j.

d: If $X^j>0$, $X^{j+1}<\infty$, the return pattern $g^j(\tilde{X}_i)$ is proportional to $\bar{\bar{s}}_i(\tilde{X}_i; \bar{\bar{d}}_i(\hat{a}_i s_i + (1-\hat{a}_i)d_i, k_i^j), \bar{k}_i^j)$ for some $k_i^j, \bar{k}_i^j$.

Thus the return patterns $g^j(\tilde{X}_i)$ are all available. Therefore, the return pattern

$$f(X_i) = \sum_{j=0}^{\infty} g^j(\tilde{X}_i)$$

is also available.

Step 2: Let $f: \mathbb{R}_+ \to \mathbb{R}_+$ be nondecreasing, with $f(0)=0$. Then there exists a sequence $\{h^n\}$ of piecewise linear, continuous, nondecreasing functions converging to f, pointwise, such that $h^n(0)=0$ for all n.

Since f is monotone, it is of bounded variation on any bounded set. Therefore f has at most countably many discontinuities. Let $X^1, X^2, \ldots$ be an enumeration of the discontinuities of f.

For any n, define the set:

$$Y^n := \{X^1, X^2, \ldots, X^n\} \cup \left\{ \frac{j}{2^n} \,\middle|\, j=0,1,2,\ldots,n2^n \right\}$$

Let $y_0=0, y_1, y_2, \ldots, y_{m(n)}$ be an enumeration of Y^n in the natural order and define the sets:

$$Y_+^n = \left\{ y_j + \frac{y_{j+1}-y_j}{n} \,\middle|\, j=0,1,2,\ldots,m(n) \right\}$$

$$Y_-^n = \left\{ y_j \frac{y_j-y_{j-1}}{n} \,\middle|\, j=1,2,\ldots,m(n) \right\}$$

$$Z^n = Y^n \cup Y_+^n \cup Y_-^n$$

Let $z_0=0, z_1, z_2, \ldots, z_{q(n)}$ be an enumeration of Z^n in the natural order and define the function $h^n: \mathbb{R}_+ \to \mathbb{R}_+$, such that:

$$h^n(X) = \begin{cases} f(z_j) + \dfrac{X-z_j}{z_{j+1}-z_j}\left[f(z_{j+1}) - f(z_j) \right] & \text{if } X \in [z_j, z_{j+1}] \\ & j=0,1,2,\ldots,q(n)-1 \\ f(z_{q(n)}) \quad \text{if } X \geqslant z_{q(n)} & \end{cases}$$

For all n, h^n is piecewise linear, continuous, and nondecreasing, with $h^n(0)=0$. Moreover, $\lim_{n\to\infty} h^n(X) = f(X)$ for all X, as was to be shown.

The proposition follows immediately from Steps 1 and 2.

REFERENCES

Baron, D. P., "Default and the Modigliani-Miller Theorem: A Synthesis," *Amer. Econ. Rev.*, Mar. 1976, *66*, 204–12.

W. Baumol and B. Malkiel, "The Firm's Optimal Debt-Equity Combination and the Cost of Capital," *Quart. J. Econ.*, Nov. 1967, *81*, 547–78.

S. Grossman and J. E. Stiglitz, "On Stockholder Unanimity in Making Production and Financial Decisions," tech rept. no. 224, IMSSS, Stanford University, Nov. 1976.

J. Lintner, "Dividends, Earnings, Leverage, Stock Prices and the Supply of Capital to Corporations," *Rev. Econ. Statist.*, Aug. 1962, *44*, 243–69.

R. C. Merton, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *J. Finance*, May 1974, *29*, 449–70.

F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer Econ. Rev.*, June 1958, *48*, 162–97.

V. L. Smith, "Corporate Financial Theory under Uncertainty," *Quart. J. Econ.*, Aug. 1970, *84*, 451–71.

______, "Default Risk, Scale, and the

Homemade Leverage Theorem," *Amer. Econ. Rev.*, Mar. 1972, *62*, 66–76.

J. E. Stiglitz, "A Reexamination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

———, "Some Aspects of the Pure Theory of Corporate Finance: Bankruptcies and Take-overs," *Bell J. Econ.*, Autumn 1972, *3*, 458–82.

———, "The Irrelevance of Corporate Financial Policy," *Amer. Econ. Rev.*, Dec. 1974, *64*, 851–66.

The Treatment of Rents in Cost-Benefit Analysis

By EDWARD FOSTER*

Most cost-benefit analysts either implicitly or explicitly use consumers' preferences as the basis for project evaluation. The opportunity cost of a government project is measured in principle not simply by reference to the supply prices of the inputs that the project absorbs, but by reference to the value of output lost when those inputs are removed from the private sector. From this point of view, monopoly profits, taxes, or external economies can raise the opportunity cost of resources, and therefore their shadow price, above their market price. Conversely, involuntary unemployment, government subsidies, or external diseconomies can lower the opportunity cost of resources, and therefore their shadow prices, below their market prices.[1] Explicit statements of the principle (and of any principle in cost-benefit analysis) are hard to come by, but E. J. Mishan gives a characteristically careful summary; Arnold Harberger's "three basic postulates for applied welfare economics" are consistent with the principle though somewhat broader; and many writers use the principle in their discussion of specific issues.[2]

Externalities aside, the phenomena that give rise to a gap between the supply price of resources and the value to consumers of what they produce generate flows of income; I shall refer to these income flows as rents, whether received by private agents or public agencies; so rents in this context could refer to monopoly profits, tax receipts, or wages that are higher than the supply price of labor, for example. Straightforward application of the principle described above leads to the conclusion that the opportunity cost of a factor of production being withdrawn from the private sector exceeds its market price to the extent that it generates rents, so it should be valued at a shadow price that reflects the rents, for cost-benefit analysis. In Section I of the paper, I demonstrate this result.

In Section II, I question its relevance. A few economists have looked at the implications of assuming that economic agents will compete for the privilege of receiving rents, and will use real resources in the process. Gordon Tullock provided a provocative discussion of the cost of rent seeking by monopolists, recently extended by Richard Posner; Harberger (ch. 7) and Michael Todaro both looked at the rent generated by a gap between urban and rural wages for unskilled workers, and Anne Krueger generalized their insights, showing the underlying identity of a whole class of rent-seeking phenomena and arguing their empirical importance. The main conclusion of this literature is that, because of the waste of resources used to compete for rents, the welfare cost of a distortion that creates rents may be far greater than what is customarily measured as deadweight loss. Here I demonstrate the obverse: the opportunity cost of a factor that generates rent in its present occupation may be lower than it at first appears, because withdrawing the factor from its present occupation may extinguish rents and thereby free some heretofore wasted resources.

I examine a polar case in which a simple, strong result obtains: Suppose that competition for rents is so intense that there is no "profit" to rent seeking because payment of market wages to the factors of production

*University of Minnesota. Helpful comments came from several colleagues, especially from Herbert Mohring and other participants in the Applied Microeconomics Workshop at the University of Minnesota, from George Borts, and a referee of this journal.

[1] This approach is sometimes rejected for commodity taxes and subsidies on grounds that the government usually imposes them for a reason: to offset externalities, or for sumptuary purposes, for example. See I. M. D. Little and James Mirrlees (sec. 12.5).

[2] See, for example, Partha Dasgupta, Stephen Marglin and Amartya Sen (pp. 50–51, 63), Harberger (pp. 54ff), Roland McKean (pp. 38ff), and Alan Prest and Ralph Turvey (pp. 692-94).

devoted to rent seeking fully absorb the rents. Then goods produced in the private sector absorb resources equal in value to their market price, part for production and part for capturing the rent. In those circumstances, resources for which a government project must pay \$1 would, if left in the private sector, have produced goods worth exactly \$1, taking account of the resources absorbed in rent seeking; the shadow wage for the resources should equal the market wage.

The result summarized above arises from a special assumption that factor supplies are fixed (and their market returns are excluded from what I have defined as rent). If I were to assume upward-sloping factor supply curves, the rents accruing to inframarginal factor units would normally accrue only to the factor owner, with no expenditure of extra effort and with no opportunity for others to bid that rent away. Such rents would not be covered by my discussion in Section II.[3]

I. The Model without Rent Seeking

This model shows that, without rent-seeking behavior, the opportunity cost for an input exceeds its market price to reflect the rent it generates. First I describe the economy and define costs and benefits for a government project.

A. *The Economy*

I shall use the following symbols:

$Q=(q_j)=n$-vector of private consumption goods

$P=(p_j)=n$-vector of market prices for private consumption goods

$X=(x_i)=m$-vector of privately owned factors of production, assumed fixed in supply

$W=(w_i)=m$-vector of market prices for factors

$F=(f_i)=m$-vector of factors used to produce private output

$G=(g_i)=m$-vector of factors used to produce government services

$Y=(y_i)=m$-vector of factors used to seek rents

$A=(a_{ij})=m\times n$ matrix of average input-output coefficients; the a_{ij} may vary both with relative factor prices and with scale of output

$K=((k_j))=n$th order diagonal matrix of average markup ratios of price to average cost; K is defined by $P=KA'W$, where the prime denotes the matrix transpose

$R=(r_j)=n$-vector of rents, defined as price minus average cost, $R=P-A'W=(K-I)A'W$

The economy consists of a single consumer who behaves as a competitive buyer. For the purchase of private consumption goods Q, he faces prices P which are not necessarily equal to marginal production costs. In addition he is supplied with a vector of government services produced by a vector of inputs G, financed by a lump sum tax. Government services may enter utility directly or they may affect utility indirectly by changing the cost of production in the private sector, or both.

The consumer supplies a fixed vector X of factors of production for which he receives a price vector W. Let A be the average input-output matrix and $F\leqslant X$ be the vector of factors required to produce private consumption goods, $F=AQ$. The matrix A is not assumed to be constant; it may depend both on relative factor prices and on scale of output. Average costs of production for each industry are given by the vector $A'W$, where the prime denotes the matrix transpose, and $R=P-A'W$, assumed nonnegative, is the vector of rents. (Note that I exclude the market-determined returns to factors of production from this definition of rents, even though the factors are assumed to be fixed in supply; this usage violates custom.)

A government project is represented by an increase in G from G^1 to G^2, with an offsetting withdrawal of factors from the private sector. The required fall in the consumption of private goods is accomplished by a rise in the lump sum tax. Private consumption falls from Q^1 to Q^2 as a result of

[3] I am grateful to a referee for this observation.

the tax; in the absence of rent seeking, factors saved by this reduction are given by

$$(1)\qquad F^1 - F^2 = A^1Q^1 - A^2Q^2$$

I shall assume that the factors required for the government project equal the factors released from the private sector:

$$(2)\qquad G^2 - G^1 = F^1 - F^2$$

B. *The Cost-Benefit Criterion*

Cost and benefit of the project may be defined in terms of the consumer's expenditure function $e(u, P, G)$. The value of e is the lowest expenditure required to attain utility level $u = u(Q, G)$ for specific values of P and G (for economy of notation, I let G, the vector of inputs used by the government, serve as an indicator of the direct government services provided to the consumer, if any). I assume that the consumer is in competitive equilibrium, so that his actual expenditure is as shown by the expenditure function; then where superscript values 1 and 2 denote values before and after the project, respectively, we have

$$e^t = e(u^t, P^t, G^t) = P^t \cdot Q^t \qquad t = 1,2$$

Define $e^3 = e(u^1, P^2, G^2)$ where e^3 is the expenditure needed to attain the initial utility level after the government project. Now we can measure the cost of the project by $c = e^1 - e^2$ (the reduction in private spending required to finance the project), and the benefit by $b = e^1 - e^3$ (the largest reduction in private spending that the consumer would accept in order to finance the project).

PROPOSITION 1: *Given the assumption that the consumer is in competitive equilibrium, the net benefit of the project $b - c$ is equal to the consumer's compensating variation in income from the project* (CV); *and*[4]

$$b - c \gtreqless 0 \textit{ according as } u^2 - u^1 \gtreqless 0$$

PROOF:

Immediate from the definition of CV, since

$$b - c = e(u^2, P^2, G^2) - e(u^1, P^2, G^2)$$

C. *The Role of Rents*

The project cost $e^1 - e^2$ may be written

$$(3)\qquad c = P^1 \cdot Q^1 - P^2 \cdot Q^2$$

This is the reduction in private consumption expenditure resulting from the project; it is not in general equal to the project's accounting cost which, evaluated at *ex post* prices, may be written

$$W^2 \cdot (G^2 - G^1) = W^2 \cdot (A^1Q^1 - A^2Q^2)$$

using (1) and (2). If price equals marginal cost equals average cost both before and after the government project, and if $W^1 = W^2 = W$, $P^t = A^{t\prime}W$, $t = 1,2$; the accounting cost of the project (inputs used, valued at market prices) then equals the opportunity cost (the value of consumer goods given up, measured at market prices).

While the opportunity cost of a project can be measured directly from (3), information coming to the cost-benefit analyst usually focusses on the inputs required for the project rather than on the outputs to be foregone, and it is customary to estimate cost by reference to the inputs. This can be done by making two corrections to the in-

[4] The equivalent variation in income, $EV = e(u^2, P^1, G^1) - e(u^1, P^1, G^1)$, gives an alternative measure of net benefit which, like CV, has the same sign as $(u^2 - u^1)$. EV has an additional virtue not shared by CV: it gives an ordinal measure of utility change, so it can be used to compare projects; see John Hause (who also gives a careful explanation of the definition of CV by means of the expenditure function). CV seems to be more widely used in the literature of cost-benefit analysis, so I have used it here; but the project cost would be defined in the same way under either alternative, so the subsequent discussion would be unaffected. In the text I have not explicitly treated possible external effects of the project, but they can easily be added: define a vector of environmental variables E and write $e^t = e(U^t, P^t, G^t, E^t)$, $t = 1,2$. Definitions of cost and benefit generalize in an obvious way, and the subsequent discussion of project cost is again unaffected because external effects are treated as (positive or negative) benefits and do not affect the cost.

puts' market value: the first adjusts for the income effect of changes in factor prices occasioned by the project, and the second adjusts for rents. We need

$$(4) \qquad P^t\cdot Q^t = W^t\cdot F^t + R^t\cdot Q^t \qquad t=1,2$$

where R is a vector of rents. Equation (4) simply specifies that revenue is divided into factor costs and rents.

PROPOSITION 2: *Given an economy in which equations* (2), (3) *and* (4) *hold, the opportunity cost for a government project may be measured as the sum of three components*:

(a) *the accounting cost of the project*, $W^2\cdot(G^2-G^1)$;

(b) *the loss of income suffered by factors originally working in the private sector,*[5] *due to the change of factor prices*, $(W^1-W^2)\cdot F^1$;

(c) *the loss of rents occasioned by the project*, $R^1\cdot Q^1 - R^2\cdot Q^2$.

PROOF:

From (3), using (2) and (4), we have

$$(5) \quad c = P^1\cdot Q^1 - P^2\cdot Q^2 = W^2\cdot(G^2-G^1) + (W^1-W^2)\cdot F^1 + (R^1\cdot Q^1 - R^2\cdot Q^2)$$

Remark: In the case where factor prices are unaffected by the project in question, $W^1=W^2=W$ and it is possible to define a vector of shadow factor prices W^* such that

$$(6) \qquad c = W^*\cdot(G^2-G^1)$$

To see how W^* may be determined, define $K=((k_i))$ as the diagonal matrix showing the ratio of goods price to average cost for each sector, and suppose that K is not affected by the project. That is, suppose

$$(7) \qquad P^t = KA^tW \qquad t=1,2$$

Then assuming that cost can be written as in (6), we derive from (1), (2), (3), (6), and (7) an equation relating W^* to W:[6]

$$(8) \qquad W^*\cdot(A^1Q^1 - A^2Q^2) = W\cdot(A^1KQ^1 - A^2KQ^2)$$

This is a single equation in m variables and therefore is not sufficient to define the individual shadow factor prices w_i^*. However each side of (8) is an inner product of m-vectors. We may derive appropriate (though not uniquely appropriate) values for the shadow prices w_i^* if we force equality of those inner products term by term. This procedure yields

$$(9) \quad w_i^* = w_i \sum_j b_{ij}k_j \Big/ \sum_j b_{ij} \qquad i=1,\dots,m$$

where

$$(10) \qquad b_{ij} = a^1_{ij}q^1_j - a^2_{ij}q^2_j$$

Equations (9) should be understood to be omitted for any factor i for which $\sum_j b_{ij}=0$; w_i^* is not then determined (but neither is it needed since the project does not use x_i).

Equation (9) says that the ratio of the shadow price to the market price for factor i is a weighted average of the k_j (price/cost ratio for sector j); the weight on k_j measures the share of factor i released to the project by sector j, as total private spending is reduced to accommodate the increased public activity.

Example: If, for each ten workers hired by the government project, four come from a sector in which the value of their marginal product (*VMP*) was 1.5 times the wage and six come from a sector in which *VMP* was twice the wage, then the shadow wage should be 1.8 times the market wage:

$$w^*/w = (4/10)\times 1.5 + (6/10)\times 2 = 1.8$$

[5] What about the income loss for factors originally employed in the public sector? Those are irrelevant, and do not appear in the calculation because such income changes are offset by opposite changes in the lump sum tax. Note that the adjustment for income changes takes the particularly simple form of (change in factor prices)×(quantity) because I have assumed that factor supplies are inelastic.

[6] When first introduced in connection with equation (2), the requirement that resources released from the private sector equal resources absorbed by the government project appeared innocuous. But if there is more than one factor of production, a great coincidence must occur for the private sector to release resources in exactly the proportions needed by the government project, with no change in factor prices. Though unrealistic, the assumption is common in cost-benefit analysis.

II. Rent Seeking

Two general mechanisms have been suggested whereby rent seeking uses resources (see Krueger). In the first, each unit of rent goes to the agent who devotes the most resources to acquiring it; in the second, rents are distributed randomly among those who compete for them, with the expected value of rent to be obtained by any agent proportional to the resources he devotes to rent seeking. Examples of the first mechanism are the allocation of sales or import licenses in proportion to the productive capacity of the license seeker (excess capacity represents the rent-seeking resource use); the award of jobs which provide rent[7] to the candidates with the best educational background (overeducation represents the rent-seeking resource use); and any of a number of practices of monopolies to exclude entry. Examples of the second, "random," mechanism are the allocation of urban jobs among unemployed rural emigrants in the Todaro model, or the use of lobbying to influence the allocation of tax revenues to specific expenditure programs.[8]

Under the first "winner take all" mechanism, there will be continuing incentive for new candidates to enter and capture the rents by spending more, so long as rents exceed the market price for the factors needed to capture those rents. Under the second random mechanism, there will be a similar continuing incentive for new candidates to enter so long as the expected return exceeds the opportunity cost, provided that rent seekers are risk neutral.

I shall use the term *competition* in rent seeking to mean a situation in which the market value of resources devoted to rent seeking equals the total value of the rents (with an appropriate adjustment for a risk premium in the case of risk aversion, considered below). In this section, I consider the implications for measurement of project costs of this extreme form of rent-seeking behavior.

A. *Risk Neutrality*

In this part I consider the effect of rent-seeking behavior on the economy described in Proposition 2, assuming risk-neutral rent seekers. Equations (3) and (4), assumed to hold in the earlier discussion, still hold; equation (2) specifying the full-employment condition is changed to

$$G^2-G^1=(F^1+Y^1)-(F^2+Y^2) \tag{11}$$

where Y^t represents the vector of factors used for rent-seeking in situation t. Equation (11) says that the factors absorbed by the government are those released by the private sector. I shall further assume that rent seekers are risk neutral. Under that assumption I have defined competitive rent seeking to mean that

$$W^t\cdot Y^t=R^t\cdot Q^t \quad t=1,2 \tag{12}$$

where R is the vector of unit rents (price minus average cost). Equation (12) says that rents are fully absorbed by payments to factors engaged in rent seeking.

PROPOSITION 3: *Given an economy in which equations* (3), (4), (11) *and* (12) *hold, the opportunity cost for a government project may be measured as the sum of two components:*

(a) *the accounting cost of the project,* $W^2\cdot(G^2-G^1)$;

(b) *the loss of income suffered by factors originally working in the private sector, due to the wage change,* $(W^1-W^2)\cdot(F^1+Y^1)$.

No adjustment to the accounting cost is required for the change in rents occasioned by the project, in contrast to the result in Proposition 2.

[7] Rent in this case could represent an excessive salary or an opportunity to receive bribes, thus sharing in the rents generated through government activity; see Krueger, pp. 292-93.

[8] This example motivated my labeling tax revenues as rents, contrary to custom. A disturbing extension of the example suggests that *all* cost-benefit analysis of government projects might be misleading: if alternative mooted projects are the object of resource-using lobbying activities, the lobbying could under varying assumptions use up part, all, or more than the total social benefit to be obtained from the selected project.

PROOF:

Directly from (3), using (4), (11), and (12), with a little manipulation, we have

$$c = W^2\cdot(G^2 - G^1) + (W^1 - W^2)\cdot(F^1 + Y^1) \tag{13}$$

In comparing this result with that of Proposition 2, keep in mind that, for economy of notation, identical symbols for factor prices and quantities are being used to describe two different regimes in the two propositions. Both factor prices and total private sector employment of all factors would have to be identical in the two regimes to claim that the first two cost components in (5) are equal to the two cost components in (13), so we cannot in general assert that the project cost is lower by the amount of the rent adjustment when there is rent seeking; but such an assertion is not relevant to my purpose. My interest is in the procedure to be used in estimating project costs, and Proposition 3 shows that the additional step of estimating the change in rents, as part of the measure of cost, is inappropriate if there is competitive rent seeking.

Note that if factor prices are constant, (13) simplifies to

$$c = W\cdot(G^2 - G^1) \tag{14}$$

The opportunity cost of the project then equals the accounting cost, and shadow prices for factors equal their market prices.

B. *Risk Aversion*

Risk neutrality by rent seekers is not essential to the argument. Suppose that rents are assigned exclusively by the random mechanism, and suppose that agents engaged in rent seeking are indifferent between receiving, for each unit of factor i, a wage of w_i with certainty and expected return of $w_i + v_i$ through the gamble of rent seeking. Then where $V = (v_i)$, we must substitute for (12) the equilibrium condition

$$(W^t + V^t)\cdot Y^t = R^t\cdot Q^t \tag{15}$$

When private output is reduced from Q^1 to Q^2, rent seeking will be reduced, but part of the loss in rent is offset by a gain that comes through reduced uncertainty. This gain could be treated as an additional benefit to the project, but it is useful instead to consider it as a reduction in cost. Under the assumption that the value of V is not affected by the reduction in the scale of rent seeking or by the government project itself, the compensating variation in income for the reduced uncertainty is $V\cdot(Y^1 - Y^2)$; that is, this measures the reduction in expected gains that rent seekers will willingly trade for reduced uncertainty.[9] With V constant, then, and using the convention that the benefit of risk reduction is to be recorded as a reduction in cost rather than an increase in benefit, we substitute for equation (3)

$$c = P^1\cdot Q^1 - P^2\cdot Q^2 - V\cdot(Y^1 - Y^2) \tag{16}$$

PROPOSITION 4: *Given an economy in which equations* (4), (11), (15), *and* (16) *hold, project cost is measured by equation* (13), *just as in the case of risk neutrality.*

PROOF:

Substituting (4) and (15) into (16) we have

$$c = W^1\cdot(F^1 + Y^1) - W^2\cdot(F^2 + Y^2)$$

Then using (11), this can be rewritten as in (13).

Except for possible complications caused by the scale of the project changing the degree of risk aversion in the economy, the basic result still holds: with competitive rent seeking, the opportunity cost of the project should not include an adjustment for rents.

[9] This discussion applies strictly only to the case of a single factor owner. If there are many agents seeking rents with different degrees of risk aversion, equation (15) would still describe equilibrium but the aggregate compensating variation in income from risk reduction would not necessarily be equal to $V\cdot(Y^1 - Y^2)$. A more detailed specification of the economy would be required for analysis of that case.

III. Summary and Conclusion

In the context of a government project that diverts inputs from the private sector, I have given a formal justification for the practice of using shadow prices for factors that differ from market prices in order to reflect the value to the consumer of private sector output foregone; I have also provided the obvious extension of this analysis to a case in which factor prices change as a result of the government project. I have then shown that the justification given depends for its validity on an assumption that no resources are devoted to rent seeking.

I define rent as the difference between factor cost and market price for private sector outputs. I define competitive rent seeking as a situation in which the market-determined prices of factors of production devoted to rent seeking completely exhaust the value of the rents (possibly discounted for the uncertainty of their receipt). When there is competitive rent seeking, there should be no correction to the factor cost of a project to account for changes in rents. The reason is that the value of resources devoted to production plus rent seeking equals the value of the goods produced, so that each dollar of foregone consumption frees one dollar of resources. While I have not shown the algebra, it is clear that intermediate cases—between competitive rent seeking and no rent seeking—lead to intermediate shadow prices for factors of production.

It seems to me that there is little question that rent seeking is important in real economies. Krueger has made a persuasive case for the pervasiveness of the phenomenon, and although she did not explicitly consider tax receipts as rents, when one considers the extent of lobbying (for example, over competing uses for the highway trust fund) and efforts devoted to capturing federal grants and contracts (for example, by universities) it is clear that at least some tax receipts are fought over.

These observations do not justify a claim that real world rent seeking is competitive. First, there are the rents that accrue to inframarginal factor units when factor supplies slope up, which are not treated in this paper but which are almost certainly not subject to competition. For other rents, however, the claim of competition fits very comfortably into a theory that assumes individual rationality, and we should at least ask "if rent seeking is not competitive, why not?" Barriers to entry into rent seeking would not provide a satisfactory answer, since it simply pushes the question one step further back (the hypothetical barriers would themselves create a second-order rent-seeking opportunity). Lack of information on the availability of rents might well offer a formal justification for noncompetitive rent seeking that is compatible with individual rationality, but it would require a more elaborate model than the one used here to explore the issue.

I conclude that for rents of the kind treated here, the prescription based on competitive rent seeking, to ignore rents in calculating project costs, may or may not be correct; but the alternative prescription, based on no rent-seeking activity at all, is wrong.

REFERENCES

J. M. Buchanan, "Taxation in Fiscal Exchange," *J. Publ. Econ.*, July/Aug. 1976, *6*, 17–29.

Partha Dasgupta, Stephen Marglin, and Amartya K. Sen, *Guidelines for Project Evaluation*, U.N. Industrial Development Organization, New York 1972.

Arnold C. Harberger, *Project Evaluation*, Chicago 1974.

J. C. Hause, "The Theory of Welfare Cost Measurements," *J. Polit. Econ.*, Dec. 1975, *83*, 1145–82.

A. O. Krueger, "The Political Economy of the Rent-Seeking Society," *Amer. Econ. Rev.*, June 1974, *64*, 291–303.

I. M. D. Little and James A. Mirrlees, *Project Appraisal and Planning for Developing Countries*, New York 1974.

R. N. McKean, "The Use of Shadow Prices," in Samuel B. Chase, ed., *Problems in Public Expenditure Analysis*, Washington 1968.

E. J. Mishan, "The ABC of Cost-Benefit," *Lloyds Bank Rev.*, July 1971, No. 101, 12–25.

R. A. Posner, "The Social Costs of Monoply and Regulation," *J. Polit. Econ.*, Aug. 1975, *83*, 807–27.

A. R. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," *Econ. J.*, Dec. 1965, *75*, 683–735.

M. P. Todaro, "A Model of Labor Migration and Urban Unemployment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, *59*, 138–48.

G. Tullock, "The Welfare Costs of Tariffs, Monopolies and Theft," *Western Econ. J.*, June 1967, *5*, 224–32.

Cartel Problems: Note

By R. ROTHSCHILD*

In a paper in this *Review*, D. K. Osborne devises a "quota rule" to be followed by members of a cartel who wish to discourage cheating. Osborne's proposition is that once a joint profit-maximizing output level has been chosen, the best reply by the "loyal" members to cheating in the form of unauthorized increases in output is to increase output in the same proportion as the cheaters increase theirs. This "market share maintenance" strategy will, he argues, leave the loyal members better off than they would have been if they had taken no action, and the cheaters worse off than they were before cheating. Hence, in anticipation of such a response from loyal members, rational potential cheaters can be expected to refrain from cheating.

In a comment on Osborne's paper, William Holahan has shown that if sufficiently large differences exist between the profit functions of the loyalists and the cheaters, the application of the quota rule may leave the loyal members worse off.

The purpose of this comment is to identify precisely conditions under which the application of the quota rule will leave loyal members worse off than they would have been if they had stood pat. In contrast to Holahan, I shall show that this may occur even when all firms have identical profit functions. In particular, I shall show that the successful application of Osborne's rule depends crucially upon the number of firms in the cartel, its initial output level, and the size of the increase in the cheaters' output.

Let us assume that the cartel is comprised of N firms, $N>2$, producing a homogeneous good. Let the demand function in inverse form be

$$(1) \qquad p=f(Q) \qquad f'<0$$

Suppose that each firm produces $1/N$th of the joint profit-maximizing output Q^m, subject to a cost function which is identical for all firms. Then the profit to each at output Q^m/N with associated cost C^m is

$$(2) \qquad \frac{Q^m}{N}f(Q^m)-C^m$$

Suppose now that *one* firm (the cheater) increases its output by the amount h. If the remaining firms follow Osborne's quota rule, the consequent increase k in the loyalists' combined output must be such that

$$(3) \qquad \frac{k}{N-1}=h$$

Then $h+k=hN$, and the share of each firm in the cartel remains $1/N$th of total output. If costs to each at the new output level are C^o, then profit is

$$(4) \qquad \left(\frac{Q^m}{N}+h\right)f(Q^m+hN)-C^o$$

This is less than the profit obtained by standing pat in the face of cheating if

$$\left(\frac{Q^m}{N}+h\right)f(Q^m+hN)-C^o < \frac{Q^m}{N}f(Q^m+h)-C^m$$

or

$$(5) \qquad 1+h\frac{N}{Q^m}=1+\frac{\Delta Q^m}{Q^m} < \frac{f(Q^m+h)+\frac{N}{Q^m}(C^o-C^m)}{f(Q^m+hN)}$$

Suppose that costs are constant (i.e., $C^o=C^m$).[1] Then (5) reduces to

$$(6) \qquad 1+\frac{\Delta Q^m}{Q^m} < \frac{f(Q^m+h)}{f(Q^m+hN)}$$

*University of Lancaster, England. I thank an anonymous referee for helpful comments on an earlier draft.

[1] Since Q^m is joint profit maximizing, we may suppose $C^m < C^o$. If $C^m < C^o$, the conditions established below will hold a fortiori.

The condition on elasticity of demand at $f(Q^m+h)$ necessary for the inequality in (6) to hold is found by setting $f(Q^m+h)=P^\alpha$ and rewriting (6) as

$$1+\frac{\Delta Q^m}{Q^m}<\frac{P^\alpha}{P^\alpha-|\Delta P^\alpha|}$$

or

$$\frac{\Delta Q^m P^\alpha}{|\Delta P^\alpha|Q^m}<1+\frac{\Delta Q^m}{Q^m} \tag{7}$$

or, equivalently,

$$|\varepsilon|_L<1+h\frac{N}{Q^m} \tag{8}$$

The measure $|\varepsilon|_L$ is to be interpreted here as the elasticity of demand between P^α and $P^\alpha-|\Delta P^\alpha|$ for the output of each of the $N-1$ loyalists, given the output of the cheater. As such, it will (by familiar argument) be greater than the elasticity of demand between P^α and $P^\alpha-|\Delta P^\alpha|$ for the output of the loyalists as a group.

Similarly, we can establish a necessary condition on the elasticity of demand for the output of the cheater, $|\varepsilon|_C$, between the original price P^m and P^α, assuming that each loyalist continues to produce Q^m/N. This is simply

$$|\varepsilon|_C>1+h\frac{N}{Q^m} \tag{9}$$

It is clearly true, since the loyalists act together, that $|\varepsilon|_L$ is smaller, and $|\varepsilon|_C$ larger, as N is larger.

Without loss of generality, we can set $h=\beta(Q^m/N)$, $\beta>0$. Then the ratio of the increase $\beta(Q^m/N)$ to the cheater's initial output Q^m/N is β, and the ratio of the increase to the initial cartel output Q^m is β/N. As presented, Osborne's analysis is insensitive to the distinction between β and β/N. In particular, it does not distinguish between (i) the case where β may be small, but β/N is large (because N is small), and (ii) the case where β may be large, but β/N is small (because N is large). Since the cheater's choice of β should be made subject to considerations not only of (9), but also of (8), this distinction is of some importance.

In order to discuss each of these cases in terms of our two elasticity measures, we substitute $\beta(Q^m/N)$ for h in (8) and (9) to obtain

$$|\varepsilon|_L<1+\beta \tag{10}$$

and

$$|\varepsilon|_C>1+\beta \tag{11}$$

Case (i): Suppose β is small and β/N large. Then the impact of cheating on the loyalists will be significant. If (11) *and* (10) hold, the quota rule cannot be applied since it will leave the loyalists worse off. It follows that if the latter are bound by the quota rule, then cheating will take place.

Realistically, however, we should expect that if (10) and (11) hold, the loyal firms will choose to respond in some other way. One possibility is that they will undertake $N-1$ equal increases in output, each by an amount less than $\beta(Q^m/N)$. For any $\beta>0$ and $|\varepsilon|_L>1$, there is an r, $0<r<1$, such that[2]

$$1+\beta>|\varepsilon|_L>1+r\beta \tag{12}$$

If each of the loyalists chooses the ratio $r\beta$ in *partial* (or non-quota-preserving) retaliation, then this, as Osborne's analysis shows,[3] will produce results qualitatively similar to those which the application of the quota rule would have achieved if (10) had not held. Consequently, no cheating will take place.

The important point here is that, given sufficiently low $|\varepsilon|_L$, the quota rule will be invalid even though β is small. Nevertheless, since β/N is large, the loyalists will wish to discourage cheating and can do so by adopting a strategy of partial retaliation.

[2] Of course, if $|\varepsilon|_L<1$, then *any* retaliation along the lines discussed here is ruled out.

[3] In Osborne's analysis, the implicit assumption appears to be that partial retaliation is the answer when *both* β and β/N are large. But his argument clearly applies here, too.

Case (ii): Suppose β is large and β/N small. If (11) *and* (10) hold, the quota rule is clearly invalid even though β is large. Since, as already indicated, $|\varepsilon|_L$ at any given price will be lower and $|\varepsilon|_C$ higher, the larger is N, the conditions specified in (10) and (11) are not unrealistic in the present context and we may suppose that, given suitably large N, large β can be found to satisfy (10) and (11). Where this is so, cheating will be encouraged but the impact on the loyalists will be small.

The implications for the quota rule of large β and small β/N can be made strikingly clear with the aid of a simple example. Assume that the cartel is of the type described earlier, with each firm initially producing $1/N$th of total output. Let the cartel demand function take the form $p=A-bQ$, where A and b are positive constants, and suppose that elasticity of demand for cartel output at P^m is close to unity. Assume that costs are negligible. If N is large, so that β/N is small when $\beta=1$, then $|\varepsilon|_C>1+\beta$. A doubling of output by the cheater will have only a small impact on cartel price. But, suppose that each loyalist also chooses $\beta=1$. Then cartel output will have doubled and, since $|\varepsilon|_L<1+\beta$ at P^α, the loyalists are worse off after retaliation on the basis of the quota rule.

The remaining question is therefore whether, in this instance, partial retaliation is an alternative. It will be clear by now that β may be quite large without this necessarily having severe effects on the loyalists. All that is required is that N also be large, which is so by assumption. In this case, therefore, the loyalists as individuals may well not be concerned to retaliate, especially since retaliation is at best partial. If this is so, cheating clearly pays. The impact of such cheating on the cartel price will, of course, be small, and it is worth noting that, as N becomes large, the cheater will be placed increasingly in the position of a "price taker."

Two conclusions to emerge from the foregoing analysis are (i) that there can be found conditions on $|\varepsilon|_L$ and $|\varepsilon|_C$ which invalidate the quota rule, and (ii) that the possibility that these conditions might be met in practice is obscured in Osborne's analysis because the absolute and relative magnitudes of β and N necessary to yield his results are not clearly specified. As presented, Osborne's rule would seem to suit best the case where market elasticity of demand is high and both β and β/N are small. Whether or not this is a condition likely to be met in practice is an essentially empirical matter.

REFERENCES

W. L. Holahan, "Cartel Problems: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 942–46.

D. K. Osborne, "Cartel Problems," *Amer. Econ. Rev.*, Dec. 1976, *66*, 835–44.

A Risk-Return Model with Risk and Return Measured as Deviations from a Target Return

By DUNCAN M. HOLTHAUSEN*

Two-attribute risk and return models are very popular in the economics and finance literature for analyzing decisions under uncertainty. Their popularity stems primarily from the intuitive appeal of the dichotomy into risk and return, and from the ease with which the concepts can be diagrammed in two dimensions.

The most commonly used risk-return model is the mean-variance model in which risk is measured as the variance and return by the mean of the probability distribution over outcomes. The mean-variance model has a number of shortcomings which are widely known, but of particular concern here are the facts that 1) mean-variance dominance is neither necessary nor sufficient for second-degree stochastic dominance; 2) unless the form of the probability distribution is restricted, mean-variance is consistent with von Neumann-Morgenstern utility theory only if the utility function is quadratic; and 3) as Peter Fishburn has noted,

> ...decision makers in investment contexts very frequently associate risk with failure to attain a target return. To the extent that this contention is correct, it casts serious doubt on variance—or, for that matter, on any measure of dispersion taken with respect to a parameter (for example, mean) which changes from distribution to distribution—as a suitable measure of risk.
> [pp. 117–18]

Recently, Fishburn has presented a risk-return model in which risk is associated only with outcomes falling below some specified target level, t. In its most general form, risk is defined by

$$(1)\qquad \rho(F)=\int_{-\infty}^{t}\phi(t-x)\,dF(x)$$

where $\phi(y)$, for $y\geqslant 0$, is a nonnegative nondecreasing function in y with $\phi(0)=0$, and $F(x)$ is the probability distribution function over outcomes x, i.e., $F(x)$ gives the probability of getting an outcome less than or equal to x. In this paper, as in Fishburn's, integrals are to be interpreted as Lebesque-Stieltjes integrals, and all probability distributions are assumed to be bounded in the sense that $F(x)=0$ and $F(y)=1$ for some real x and y. This simplifies the mathematics with no loss in reality.

In Fishburn's model, the distribution F dominates another distribution G if and only if $\mu(F)\geqslant\mu(G)$ and $\rho(F)\leqslant\rho(G)$ with at least one strict inequality, where $\mu(F)$ is the mean of F. A specific form of (1) is the α-t model which is much simpler to estimate. In the α-t model, risk is defined by

$$(2)\qquad r(F)=\int_{-\infty}^{t}(t-x)^{\alpha}d(F)\qquad \alpha>0$$

Using the α-t model, Fishburn has shown congruence between that model and an expected utility model in which the utility function is

$$(3)\qquad \begin{aligned} U(x)&=x &&\text{for all } x\geqslant t\\ U(x)&=x-k(t-x)^{\alpha} &&\text{for all } x\leqslant t \end{aligned}$$

where k is a positive constant. The decision maker may display various degrees of risk aversion or preference for outcomes below t depending on the value of α, but he is risk neutral for outcomes above t. After surveying a number of empirical studies of utility functions, Fishburn concludes

> ...that most individuals in investment contexts do indeed exhibit a target return—which can be above, at, or below the point of no gain and no loss—at which there is a pronounced change in the shape of their utility functions,

*Professor, department of economics and business, North Carolina State University.

> and that [the utility function given in (3)] can give a reasonably good fit to most of these curves in the below-target region. However, the linearity of [(3)] holds only in a limited number of cases for returns above target.
>
> [p. 122]

The purpose of this paper is to present a risk-return model that has many of the same attributes as Fishburn's model, but one in which the utility function for above-target outcomes need not be linear.

I. The Model

Risk is defined as in Fishburn's paper by equation (1) in the general case and by equation (2) as a specific example. Return, however, is measured by a probability-weighted function of deviations above the target level t. The general form of the return measure is

$$\Pi(F)=\int_t^{\infty}\Psi(x-t)dF(x) \tag{4}$$

where $\Psi(y)$, for $y\geqslant 0$, is a nonnegative nondecreasing function in y with $\Psi(0)=0$. A distribution F will be said to dominate a distribution G if and only if $\Pi(F)\geqslant\Pi(G)$ and $\rho(F)\leqslant\rho(G)$ with at least one strict inequality. Let us call this the ρ-Π-t model. The measure of return used here is different from others given in the literature, because it only depends on outcomes above the target level. In many ways, this is intuitively more appealing than using a measure, like the mean, which is based on all possible outcomes above and below the target. Since outcomes below the target outcome have already been considered in the risk measure ρ, it seems redundant if not contradictory to include them in the return measure. Another feature of this return measure is that it does not impose risk neutrality on the decision maker for outcomes above t.

A specific form of the risk function which permits easy estimation is

$$R(F)=\int_t^{\infty}(x-t)^{\beta}dF(x) \qquad \beta\geqslant 0 \tag{5}$$

Using (2) as the risk measure along with (5) gives an α-β-t model. The decision maker is risk averse (risk neutral) (risk seeking) for returns above t if β is less than (equal to) (greater than) one. In similar fashion, the decision maker is risk averse (risk neutral) (risk seeking) for returns below t if α is greater than (equal to) (less than) one, as Fishburn has shown. Letting $P(\alpha,\beta,t)$ denote dominance by the α-β-t model, we say $FP(\alpha,\beta,t)G$ if and only if $R(F)\geqslant R(G)$ and $r(F)\leqslant r(G)$ with at least one strict inequality.

Estimating values for α and β is a fairly straightforward process. In fact, estimation of α is much easier in the α-β-t model than it is in Fishburn's α-t model. To estimate α, consider two lotteries in which all outcomes are at or below t. Lottery one gives t with probability p and $t-2d$ with probability $(1-p)$, and lottery two gives $t-d$ with certainty. The amount $d>0$ should be a "significant" amount. Let us call these distributions $F(x)$ and $G(x)$, respectively. Then $R(F)=R(G)=0$ because no outcomes are above t. The risk measures are $r(F)=(2d)^{\alpha}(1-p)$ and $r(G)=d^{\alpha}$. If p_0 is the probability at which the decision maker is indifferent between the two lotteries, then $r(F)=r(G)$ and $\alpha=log\ (1/(1-p_0))/log\ 2$. The same method can be used to estimate β. In that case, compare the lottery with payoffs t and $t+2d$ against the sure thing $t+d$. At the point of indifference, p', $R(F)=R(G)$ and $\beta=log\ (1/(1-p'))/log\ 2$.

Since the α-β-t model is an approximation of a possibly more complex model, it may be that the estimated values of α and β are sensitive to the choice of d. The decision maker will have to decide if the approximation is "close enough." If not, either the more general model must be used, or it may be that none of the models in this paper are appropriate.

II. Congruence with Expected Utility

The ρ-Π-t dominance model is a version of a more precise preference relationship in which preference depends only on $\Pi(F)$ and $\rho(F)$. If $V(\Pi(F),\ \rho(F))$ is a real valued function which is increasing in Π and decreasing in ρ, then the decision maker's preferences satisfy the ρ-Π-t model if for all relevant distributions F and G,

(6) F is preferred to G iff

$$V(\Pi(F),\rho(F)) > V(\Pi(G),\rho(G))$$

Now, assuming such a function V exists, it is not necessary that the decision maker's preferences also satisfy the von Neumann-Morgenstern axioms for expected utility. What we would therefore like to show is that there is a von Neumann-Morgenstern utility function $U(x)$, such that

$$(7)\quad V(\Pi(F),\rho(F)) > V(\Pi(G),\rho(G))$$

$$\text{iff}\quad \int_{-\infty}^{\infty} U(x)\,dF(x) > \int_{-\infty}^{\infty} U(x)\,dG(x)$$

THEOREM 1: *Suppose that, for all bounded distribution functions F and G, the* ρ-Π-t *model is congruent with the expected utility model in the sense that* (7) *holds*; *then letting* $U(t)=0$ *and* $U(t+1)=1$, *there exists a positive constant k such that*

$$(8)\quad U(x)=\begin{cases} \Psi(x-t)/\Psi(1) & \text{for all } x \geqslant t \\ -k\phi(t-x) & \text{for all } x \leqslant t \end{cases}$$

(The proof is given in the Appendix.)

What the theorem shows is that there is an expected utility function which is congruent with the ρ-Π-t model, so this risk-return model is consistent with the von Neumann-Morgenstern axioms, and preferences accordant with the model can be represented by a von Neumann-Morgenstern utility function.

Taking expected utility of (8) gives

$$\int_{-\infty}^{\infty} U(x)\,dF(x) = [1/\Psi(1)]\Pi(F) - k\rho(F)$$

so indifference curves in risk and return space are parallel straight lines. When the α-β-t model is used, the specific form of the congruent utility function is

$$(9)\quad U(x)=\begin{cases} (x-t)^{\beta} & \text{for all } x \geqslant t \\ -k(t-x)^{\alpha} & \text{for all } x \leqslant t \end{cases}$$

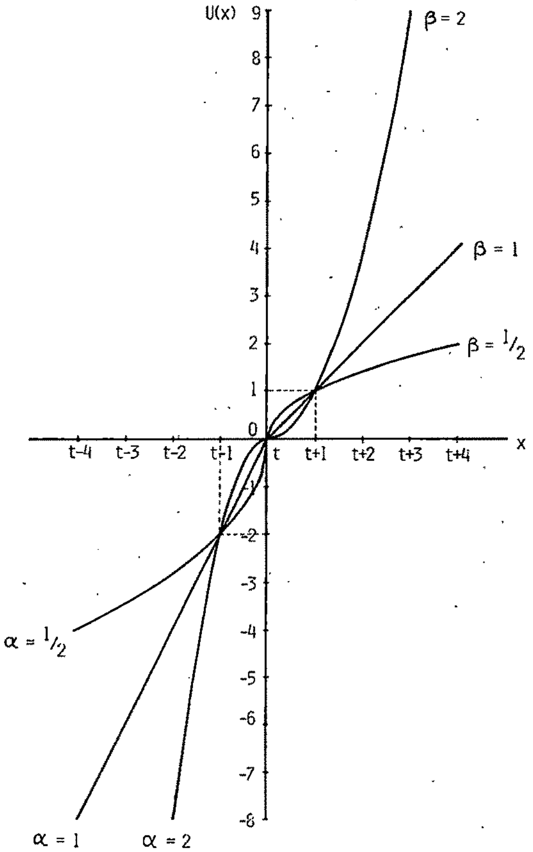


FIGURE 1. PLOTS OF THE UTILITY FUNCTION GIVEN BY (9) FOR $k=2$ AND VARIOUS α, β VALUES

The solution for k is then quite straightforward since $U(t-1)=-k$.

A utility function of the type given in (9) can take on a number of shapes depending on the values of α, β, and k. Figure 1 depicts some possibilities when $k=2$. Various risk profiles can be generated by choosing appropriate values of α and β. For example, if $\alpha<1$ and $\beta<1$, then the individual is risk average above t and risk seeking below t. If the individual is risk neutral above and below t (i.e., $\alpha=\beta=1$), he still displays a form of risk aversion for lotteries that have payoffs above and below t because there will be a kink at t as long as $k>1$. If $\alpha=\beta=1$ and $k=1$, the individual is completely risk neutral, and if $k<1$ while

$\alpha=\beta=1$, he displays a form of risk preference for lotteries spanning t. Finally, if the individual is risk averse above and below t (i.e., $\alpha>1$ and $\beta<1$), it is possible that he will be risk seeking in a small neighborhood around t because of the kink at t. This would be the case in Figure 1 if $\alpha=2$ and $\beta=1/2$. Such a utility function is a form of the Friedman-Savage utility function and is constant with the oft noted fact that many people gamble small amounts at fair or slightly unfair odds and yet, at the same time, buy insurance against large risks.

Reviewing some of the empirical literature in which utility functions have been estimated, it appears that the α-β-t model fits almost all of the functions reasonably well. Ralph Swalm gives estimated utility functions for thirteen corporate executives. In twelve of the cases, there is a point at which there is a pronounced change in the shape of the utility function. This point is approximately zero (the break-even point on a project) in each case. The concept of a target level of return thus seems quite appropriate for these decision makers. In the majority of cases, the executives are risk seekers below zero and risk averse above. Therefore, α and β are generally less than one. Utility also tends to drop sharply below t indicating fairly large values of k. There are a few executives with linear or risk-averse shapes below t and two who are risk seekers above t, but these are all consistent with various forms of the α-β-t model.

Jackson Grayson's study of oil drilling decision makers reported by Albert Halter and Gerald Dean (p. 204) also shows utility functions fairly consistent with the α-β-t model. Three of the four functions display risk-seeking behavior above t and one is risk averse. Below t, most are slightly risk averse or risk neutral and all drop steeply indicating large k values.

The utility functions in Paul Green's study of four middle managers in a large chemical company also show a good fit with the α-β-t model. The target level is about a 20 percent return on investment for each manager. All have approximately linear utility above t, so $\beta\approx1$. All are also risk averse below t, indicating that $\alpha>1$.

Finally, Halter and Dean (p. 64) show utility functions for changes in net wealth for an orchard farmer, a grain farmer, and a college professor. All have function changes at $t=0$. The orchard farmer is risk neutral above and below t (so $\alpha=\beta=1$), but slope changes significantly at t. My estimate of k is about 4.0. The college professor is risk averse above t and slightly risk seeking below.[1] My estimates of α, β, and k are .75, .62, and 3.4. The grain farmer's function does not fit as well as the others above t, because there are two concave regions separated by a convex region. My rough estimate of β is .8. Below t he becomes a strong risk seeker with $\alpha\approx.40$ and $k\approx15.8$.[2]

This review of empirically estimated utility functions indicates that the α-β-t model does capture many of the features of actual utility functions. In particular, almost all functions have a point at which their shapes change markedly. This target return point can be positive or negative and is often zero. The functions surveyed displayed many combinations of risk-averse, risk-neutral, and risk-seeking behavior. Almost all could be expressed in the α-β-t form. In this respect, the α-β-t model seems quite general.[3]

[1] Halter and Dean suggest the following explanation for such risk preferences: "When confronted with the S-shaped portion of the utility function (in the loss region), the professor explained that in his financial position losses of $20,000 to $30,000 would be quite disastrous, and that larger losses would not be viewed as proportionately more serious" (p. 63).

[2] It should be emphasized that the shapes of the utility functions and the implied risk preferences above and below the target return for each of the individuals mentioned here are not peculiar to the formulation of the α-β-t model. These are the shapes found by the researchers cited. I have only tried to fit these observed functions with the α-β-t model.

[3] Fishburn and Gary Kochenberger have recently published research that supports this conclusion. They fitted two-piece utility functions to data contained in many of the empirical utility studies mentioned above. They found that power functions of the form given in (9) fit better, both above and below t, than either linear or exponential functions in most cases. In particular, they found that a majority of the utility functions were risk seeking below t and risk averse above. These were best fit by power functions. Exponential functions gave better fits for the more unusual case of risk-averse preferences below t and risk seeking above.

III. Stochastic Dominance Results

Stochastic dominance criteria have recently become popular because they are consistent with von Neumann-Morgenstern utility theory and avoid some of the problems associated with the mean-variance dominance model. The three major stochastic dominance criteria (first, second, and third-degree stochastic dominance) are defined as follows:

F *FSD* G iff $F \neq G$ and $F(x) \leqslant G(x)$ for all x

F *SSD* G iff $F \neq G$ and $F_1(x) \leqslant G_1(x)$ for all x

F *TSD* G iff $F \neq G$, $F_1(\infty) \leqslant G_1(\infty)$ and $F_2(x) \leqslant G_2(x)$ for all x, where

$$F_1(x) = \int_{-\infty}^{x} F(y)dy, F_2(x) = \int_{-\infty}^{x} F_1(y)dy$$

and $F_1(\infty)$ stands for $F_1(x)$ when x equals the upper limit of integration. It has been shown by Josef Hadar and William Russell, Giora Hanoch and Haim Levy, and G. A. Whitmore, among others that the three stochastic dominance criteria correspond to different classes of utility functions. Let $E(U,F) = \int_{-\infty}^{\infty} U(x)dF(x)$ be the expected utility under distribution F, then the following can be shown:

If F *FSD* G, then $E(U,F) \geqslant E(U,G)$ for all U with $U' \geqslant 0$.

If F *SSD* G, then $E(U,F) \geqslant E(U,G)$ for all U with $U' \geqslant 0$ and $U'' \leqslant 0$.

If F *TSD* G, then $E(U,F) \geqslant E(U,G)$ for all U with $U' \geqslant 0, U'' \leqslant 0$ and $U''' \geqslant 0$.

Thus, *FSD* applies to any nondecreasing utility function, *SSD* applies to all risk-averse utility functions, and *TSD* applies to all risk-averse utility functions with $U''' \geqslant 0$. Note that if a utility function is characterized by decreasing absolute risk aversion, then $U''' > 0$. So *TSD* applies to all decreasingly absolute risk-averse functions as well as others that may have $U''' \geqslant 0$.

For the α-β-t model, the following theorem is proved in the Appendix.

THEOREM 2: *If F FSD G, then F $P(\alpha,\beta,t)G$ for all $\alpha \geqslant 0$ and $\beta \geqslant 0$ except when $r(F) = r(G)$ and $R(F) = R(G)$.*

If F SSD G, then F $P(\alpha,\beta,t)G$ for all $\alpha \geqslant 1$ and $0 \leqslant \beta \leqslant 1$ except when $r(F) = r(G)$ and $R(F) = R(G)$.

If F TSD G, then F $P(\alpha,\beta,t)G$ for all $\alpha \geqslant 2$ and $0 \leqslant \beta \leqslant 1$ except when $r(F) = r(G)$ and $R(F) = R(G)$.

Thus the α-β-t efficient set is a subset of the appropriate stochastic dominance efficient set for specific values of α and β. Any distribution which is undominated by the α-β-t criterion would then also be undominated by the more general stochastic dominance criterion.

IV. Summary

A risk-return model with both risk and return measured as deviations from a target return has been developed in this paper. Risk is associated only with below-target outcomes, and return is measured only by above-target outcomes. This is apparently the way in which many decision makers view risk and return. A specific form of the model is the α-β-t model where risk is given by $r(F) = \int_{-\infty}^{t} (t-x)^{\alpha} dF(x)$ and return by $R(F) = \int_{t}^{\infty} (x-t)^{\beta} dF(x)$. The distribution F is then said to dominate another distribution G if and only if $r(F) \leqslant r(G)$ and $R(F) \geqslant R(G)$ with at least one strict inequality.

The α-β-t model can describe a variety of risk attitudes above and below the target return t, depending on the values of α and β. If α is greater than (equal to) (less than) one, the decision maker is risk averse (risk neutral) (risk seeking) for payoffs below the target return. Similarly, if β is less than (equal to) (greater than) one, the decision maker is risk averse (risk neutral) (risk seeking) for payoffs above the target.

One major result derived in the paper is that the α-β-t model is congruent with a von Neumann-Morgenstern utility function of the form

$$U(x) = \begin{cases} (x-t)^{\beta} & \text{for all } x \geqslant t \\ -k(t-x)^{\alpha} & \text{for all } x \leqslant t \end{cases}$$

where k is a positive constant and, by construction, $U(t) \equiv 0$ and $U(t+1) \equiv 1$. The magnitude of k determines the degree to which the utility function is kinked at the target return. Reviewing some empirical literature reveals the fact that most utility functions do change significantly at some point, which is often (but not always) zero. In addition, a variety of risk-averse and risk-seeking combinations can be found, all of which are fit rather well by the α-β-t model.

A second major result is that the α-β-t model is consistent with first, second, and third-degree stochastic dominance for appropriate values of α and β. That is, if distribution F dominates distribution G by one of the stochastic dominance criteria, then it will also dominate by the α-β-t criterion.

Appendix

PROOF of Theorem 1:

To simplify notation, let $t=0$. Then $U(0)=0$ and $U(1)=1$ by definition. To prove the theorem, four cases must be considered. In each case $\phi(0)$ and $\Psi(0)$ are assumed to equal zero, and $\phi(1)$ and $\Psi(1)$ are assumed positive. Finally, let $k=-U(-1)/\phi(1)$.

Case 1: $x>1$. Consider the lottery that gives x with probability $\Psi(1)/\Psi(x)$ and 0 with probability $1-(\Psi(1)/\Psi(x))$ versus the certain outcome 1. The lottery has $\Pi=\Psi(x)[\Psi(1)/\Psi(x)]=\Psi(1)$ and $\rho=0$; so does the sure thing. Therefore $V(\Pi,\rho)=V(\Psi(1),0)$ for both the lottery and the sure thing, and by (7), $[\Psi(1)/\Psi(x)]\ U(x)+[1-(\Psi(1)/\Psi(x))]\ U(0)=U(1)$. Since $U(0)=0$ and $U(1)=1$, $U(x)=\Psi(x)/\Psi(1)$. Thus (8) holds in this case.

Case 2: $0<x<1$. Compare the lottery that gives 1 with probability $\Psi(x)/\Psi(1)$ and 0 with probability $1-(\Psi(x)/\Psi(1))$ to the sure thing x. The lottery has $\Pi=\Psi(x)$ and $\rho=0$, and so does the sure thing. Thus $[\Psi(x)/\Psi(1)]\ U(1)+[1-(\Psi(x)/\Psi(1))]\ U(0)=U(x)$ and $U(x)=\Psi(x)/\Psi(1)$. Therefore (8) holds in Case 2.

Case 3: $-1<x<0$. Consider the lottery that gives -1 with probability $\phi(-x)/\phi(1)$ and 0 with probability $1-(\phi(-x)/\phi(1))$, and compare that with the sure thing providing x with certainty. $\Pi=0$ and $\rho=\phi(-x)$ in both cases. Thus $[\phi(-x)/\phi(1)]\ U(-1)+[1-(\phi(-x)/\phi(1))]\ U(0)=U(x)$, and $U(x)=U(-1)\,[\phi(-x)/\phi(1)]=-k\phi(-x)$. Hence (8) holds in this case.

Case 4: $x<-1$. Let the lottery give x with probability $\phi(1)/\phi(-x)$ and 0 with probability $1-(\phi(1)/\phi(-x))$. The sure thing gives -1 for certain. In both cases $\Pi=0$ and $\rho=\phi(1)$. Thus $[\phi(1)/\phi(-x)]\ U(x)+[1-(\phi(1)/\phi(-x))]\ U(0)=U(-1)$, which implies $U(x)=U(-1)\ [\phi(-x)/\phi(1)]=-k\phi(-x)$. Hence (8) holds in Case 4.

In all cases, the proofs do not have to be confined to outcomes only on one side of t. For example, having shown Cases 1 and 2, Case 3 can be shown to hold in the following way for lotteries that span t. Let F_1 be the fifty-fifty chance for x or $w>0$. Then let F_2 give -1 with probability $\phi(-x)/2\phi(1)$ and $z>0$ with probability $[2\phi(1)-\phi(-x)]/2\phi(1)$, where $\Psi(w)=[(2\phi(1)-\phi(-x))/\phi(1)]\Psi(z)$. The return $\Pi_1=\Psi(w)/2=[(2\phi(1)-\phi(-x))/2\phi(1)]\Psi(z)$, and thus $\Pi_1=\Pi_2$. Also $\rho_1=\rho_2=\phi(-x)/2$. Therefore, $1/2\ \ U(x)+1/2\ \ U(w)=[\phi(-x)/2\phi(1)]\ U(-1)+[(2\phi(1)-\phi(-x))/2\phi(1)]\ \ U(z)$. Solving for $U(x)$ and using the results of Cases 1 and 2 for $U(w)$ and $U(z)$ gives $U(x)=[\phi(-x)/\phi(1)]\ \ U(-1)\ -\Psi(w)+[(2\phi(1)-\phi(-x))/\phi(1)]\ \Psi(z)$. Thus $U(x)=[\phi(-x)/\phi(1)]\ \ U(-1)=-k\ \ \phi(-x)$, as I wished to show.

PROOF of Theorem 2:

If F *FSD* G, then $r(F)\leqslant r(G)$ for $\alpha\geqslant 0$ by Fishburn's Theorem 3. It therefore remains to show that $R(F)\geqslant R(G)$ for $\beta\geqslant 0$. $R(F)-R(G)=\int_t^\infty(x-t)^\beta\ [dF(x)-dG(x)]$. Integrating by parts and recalling that $F(\infty)=G(\infty)=1$, $R(F)-R(G)=-\beta\int_t^\infty(x-t)^{\beta-1}[F(x)-G(x)]\,dx$. By *FSD*, $F(x)\leqslant G(x)$ for all x, and with $\beta>0$, it follows that $R(F)\geqslant R(G)$ as was to be shown.

If F *SSD* G, then $r(F)\leqslant r(G)$ for $\alpha\geqslant 1$ by Fishburn's Theorem 3. Integrating $R(F)-R(G)$ by parts a second time yields $R(F)-R(G)=-\beta(a-t)^{\beta-1}\,[F_1(\infty)-G_1(\infty)]+\beta(\beta-1)\ \int_t^\infty(x-t)^{\beta-2}\,[F_1(x)-G_1(x)]\,dx$, where a is used to stand for the upper limit of integration. By *SSD*, $F_1(x)\leqslant G_1(x)$ for all x; thus with $0 \leqslant \beta \leqslant 1$, $R(F) \geqslant R(G)$ and $F\,P(\alpha, \beta, t)G$.

If F *TSD* G, then $r(F) \leqslant r(G)$ for $\alpha \geqslant 2$ by Fishburn's Theorem 3. Integrating $R(F) - R(G)$ by parts a third time yields $R(F) - R(G) = -\beta(a-t)^{\beta-1}[F_1(\infty) - G_1(\infty)] + \beta(\beta-1)\{(a-t)^{\beta-2} - \int_t^\infty (\beta-2)(x-t)^{\beta-3}[F_2(x) - G_2(x)]\,dx\}$. By *TSD*, $F_1(\infty) \leqslant G_1(\infty)$ and $F_2(x) \leqslant G_2(x)$ for all x. Therefore, with $0 \leqslant \beta \leqslant 1$, $R(F) \geqslant R(G)$, and thus $F\,P(\alpha, \beta, t)G$.

REFERENCES

P. C. Fishburn, "Mean-Risk Analysis with Risk Associated with Below-Target Returns," *Amer. Econ. Rev.*, Mar. 1977, *67*, 116–26.

______ **and G. A. Kochenberger,** "Two-Piece von Neumann-Morgenstern Utility Functions," *Decision Sci.*, Oct. 1979, *10*, 503–518.

M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *J. Polit. Econ.*, Aug. 1948, *56*, 279–304.

P. E. Green, "Risk Attitudes and Chemical Investment Decisions," *Chem. Eng. Progress*, Jan. 1963, *59*, 35–40.

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, *59*, 25–34.

Albert N. Halter and Gerald W. Dean, *Decisions Under Uncertainty*, Cincinnati 1971.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

R. D. Swalm, "Utility Theory—Insights into Risk Taking," *Harvard Bus. Rev.*, Nov./Dec. 1966, *47*, 123–36.

G. A. Whitmore, "Third-Degree Stochastic Dominance," *Amer. Econ. Rev.*, June 1970, *60*, 457–59.

Sweepstakes Contests: Analysis, Strategies, and Survey

By Edward B. Selby, Jr. and William Beranek*

To attract attention to product and business promotions, advertising agencies bombard the public each year with hundreds of sweepstakes contests — random drawings which differ from lotteries in that no explicit entry fee is required. Substantial participation takes place even though contest advertisements frequently lack adequate prize information, and in many cases it is difficult, if not impossible, to estimate roughly the probability of winning. Adding more complications is the existence of multiple prizes where the participants may be restricted to winning only one prize. Even when adequate information is available we find that, when consideration is given to the opportunity costs of entry, only a limited number of contests would attract the risk-neutral or risk-averse participant, in the Friedman-Savage sense, implying that most contests are not worth entering. Since these not-worth-entering sweepstakes annually draw millions of entries, must we conclude that these entrants are simply risk seekers as implied by the Friedman-Savage hypothesis, or could they be either risk neutral or risk averse as well? Participation of risk neutrals in actuarially unfair contests can be reconciled by the oft-suggested hypothesis that, in addition to wealth and risk, the individual utility function depends on "pleasures of gambling" and "occupation of idle time," nonpecuniary factors that have not been adequately explored.[1]

This paper is divided into two sections. First, entry conditions for both single and multiple-prize contests models are set forth to predict entry behavior of risk-neutral persons as well as the optimum number of entries they would submit to maximize expected profits. Based on a survey of national advertisements, the second section discusses the quality of information revealed by the advertisements, identifies the contests that a risk-neutral or risk-averse person would enter as well as those that would be avoided, determines the optimum number of entries a person would submit for each feasible contest as well as the expected return from the entries, and tests the hypothesis of risk aversion among participants. A test discriminating between the pleasures of gambling and occupation of idle time is suggested, and exploratory evidence is examined.

I. Entry Conditions

A. *Single Prize Contests*

We assume that the participant is motivated to maximize expected pecuniary gain (i.e., that he or she has a utility function which depends only on individual wealth). In decision theory this person would be described as being risk neutral, that is, would take on all actuarially fair bets. In the simplest game there is only one prize and, if all relevant information is disclosed, we set forth the following definitions:

S = dollar value of the prize
C = unit cost of participation in dollars
p = probability of winning S.

The participant either wins the prize S or loses and receives nothing. However, the unconditional cost of entry is C which includes the opportunity cost of one's time, postage, stationary, etc. Consequently the gain from winning is $S-C$ while the loss from losing is $\$0-C$. If only one entry is permitted the risk-neutral person, then it is

*University of Georgia. We are indebted to William C. Hunter and to an anonymous referee for helpful comments, and to W. Bryan Shipman, Jr. and Prabhakant Sinha for their computational assistance.

[1] For example, Donald Davidson et al. sought to neutralize these factors in order to get more accurate estimates of utility functions that had plagued the studies of Frederick Mosteller and Philip Nogee. But no effort was made to study these factors.

well known that the condition for entry is

$$p(S-C)+(1-p)(\$0-C)\geqslant \$0$$

which implies

$$pS\geqslant C \tag{1}$$

In words, the expected value of the prize must be equal to or greater than the certain cost of entry.

Suppose now the entrant can submit any number of entries, E_y. If E_0 denotes entries by all other contestants, p is now given by

$$\frac{E_y}{E_y+E_0}$$

and if entry costs are C per entry, expected winnings will be represented by

$$\frac{E_y}{E_y+E_0}(S-CE_y)$$

and expected losses by

$$\frac{E_0}{E_y+E_0}(\$0-CE_y)$$

Expected profits π become

$$\pi=\frac{E_y}{E_y+E_0}(S-CE_y)+\frac{E_0}{E_y+E_0}(\$0-CE_y)$$

or

$$\pi=\frac{E_y}{E_y+E_0}S-CE_y=E_y\left(\frac{S}{E_y+E_0}-C\right) \tag{2}$$

When the quantity in parenthesis on the right side of equation (2) is expressed in the equivalent form $[S-C(E_y+E_0)]/(E_y+E_0)$, a mildly surprising implication follows: in order for participant-expected profits to be positive, it is necessary, assuming the unit cost of all other entries is likewise C, that the value of the prize exceed the total cost of *all* entries (not just the cost for the given individual). Moreover, if risk-neutral or risk-averse individuals are to be attracted, we must have $S/C\geqslant E_y+E_0$.

Choosing E_y so as to maximize π implies satisfying the condition

$$E_y=-E_0+\sqrt{\frac{E_0S}{C}} \tag{3}$$

The reader can quickly verify that (3) yields a positive solution if, and only if, (2) is positive, and that the second-order condition is satisfied since $E_0>0$. Given E_0, S must be sufficiently large relative to C to render E_y positive; in fact, the larger the ratio S/C the greater the number of entries a contestant will submit.[2] Moreover, given the size of the prize relative to the unit cost of entry, the contestant's number of entries will be greater the *smaller* the number of all other entries, that is, the derivative of E_y with respect to E_0 is expected to be negative since the square root of E_0 is normally large relative to the square root of S/C. There has been a tendency to arrive at the opposite conclusion by focusing on the maximization of p, the probability of winning, which, at first blush, may be intuitively appealing. The reason why this logic fails, of course, is because while we increase p with E_y, total cost CE_y also increases, thus reducing the net prize. The incremental net gain declines with E_y, and this tends to swamp the benefit obtained by increasing p.

B. *Multiple-Prize Contests*

Multiple-prize contests are of two forms: the entrant is eligible to win all prizes; or

[2] Both equations (2) and (3) (as we have just noted) hold implications for advertising sponsors. To maximize both the number of individual entries as well as the number of entries per participant, they should make the ratio S/C as large as possible. Satisfying the condition $S/C>E_y+E_0$ (which makes (2) positive) will provide a necessary condition for attracting the risk averse, while at the same time being a sufficient condition for enticing the risk neutral to participate. Making S large is an obvious attraction even without the benefit of this analysis, but the benefits from minimizing entry costs may not be so obvious. Since empirical estimates of S/C are always markedly in excess of 1, a slight decrease in C is equivalent to a large increase in S. Consequently, emphasizing efforts to reduce C can often be more effective in increasing entries than a substantial increase in the size of the prize.

only one prize. Considering first the former case, if the sum of all prizes is denoted as S and since the entrant may win each prize, the analysis set forth by equations (2) and (3) above would apply.

If the entrant is eligible for only one prize, however, and there are m prizes where a typical prize is denoted by $S_i(i=1, 2,\ldots,m)$, and if each of the "other" entrants submits only one entry, the entrant who submits E_y entries will have expected winnings of

$$(4) \qquad \sum_{i=1}^{m} p_i Q_{i-1} S_i$$

where the prizes are ordered in the sequence they are drawn, Q_{i-1} is the probability of not winning on every one of the $i-1$ preceding prize drawings, i.e.,

$$Q_{i-1} = \prod_{\tau=2}^{i} \frac{E_0-(\tau-2)}{E_y+E_0-(\tau-2)} \qquad i=2,3,\ldots,m$$

$$Q_0=1,$$

and where

$$(5) \qquad p_i = \frac{E_0}{E_y+E_0-(i-1)}$$

the probability of winning the ith prize S. Because one can easily discern the relationship

$$Q_{i-1}=Q_{i-2}(1-P_{i-1})$$

Equation (4) may be rewritten in a more informative form

$$(6) \qquad \sum_{i=1}^{m} p_i \prod_{\tau=0}^{i-1} (1-P_\tau) S_i \quad P_0=0$$

Expected profit would be equation (6) less the certain cost CE_y. As the reader can quickly discern, deriving an expression for E_y comparable to (3) does not come about easily.[3] Complicating matters further is the fact that, depending on the magnitude of the prizes and the order in which they are drawn, there could be multiple local optima. Search procedures must be invoked to locate these, and even when one or more have been identified there is the uncomfortable feeling that possibly others, including the "global" optimum, are still undetected.

II. Survey Results and Empirical Tests

If sweepstakes contestants are either risk neutral or risk averse, the expected net gain from contest participation should be positive. To shed some light on this hypothesis an exploratory sample was taken of all sweepstakes contests appearing in the *Reader's Digest* during 1976 and 1977. Questionnaires were sent to the sponsors and the results are portrayed in Table 1.

Column 1 ranks the fifteen sweepstakes by the total value of prizes offered where, it must be noted, computation of this value was not always easy. Only six contests fully revealed individual prize values while in three cases total prize value could be calculated. For the remainder, estimates were necessary. The potential entrant was left to determine for himself the value of such things as an African safari or a "night on the town." In eleven of the fifteen cases, the sponsors provided the number of valid entries received (E_0). The others either failed to respond or declared the information to be confidential. No information was provided regarding the order in which prizes would be drawn. Expected winnings for a single entry in each of these multiple-prize contests differed very little, whether one assumed that the contestant could win all prizes or just a single prize. Since hindsight provided our data for E_0, calculation of expected winnings for the "all prizes" approach is shown in Table 2. The corresponding quantities for "one prize only" eligibility appears in Table 3. Comparison of Tables 2 and 3 confirm the insensitivity of expected winnings to the type of eligibility.

This insensitivity, however, does not extend to action that is required to maximize expected profits. If, for example, an allowance is made for the opportunity cost of

[3] For example, even if $m=2$ the condition for maximizing expected profits implies an expression containing a cubic form of E_y.

TABLE 1—SWEEPSTAKES SURVEYED RANKED BY VALUE OF PRIZES

				Individual Prize Values			Total Value of Prizes	
Sweepstakes	Sponsor or Coordinating Agency	Total Value of Prizes[a]	Number of Entries	Fully Revealed	Partially Revealed	Not Revealed	Given or Easily Calculated	Not Given
The Great Meow Mix Meow Off	Ralston Purina Co.	$150,000	500,000		×		×	
Glass Plus	Texize Chemicals Co.	124,048	1,200,000		×			×
G.E. New Car & Bike	General Electric Co.	72,400	350,000			×		×
Miss America	The Gillette Co.	69,000	347,166			×	×	
Miss Clairol Silver Anniversary	Clairol, Inc.	56,750	566,215	×			×	
Help Young America Save Energy	Colgate-Palmolive Co.	55,000[b]	308,000	×			×	
Help Young America	Colgate-Palmolive Co.	52,000[b]	2,044,000	×			×	
Feetstakes	Pharmacraft Consumer Products, Pennwalt Corp.	45,975	500,000		×			×
Lysol Products	Lehn & Fink Products, Co.	45,000	150,000		×		×	
Pentel	Pentel of America, Ltd.	26,700	n.a.			×		×
Vaseline Autumn	Chesebrough-Pond's, Inc.	25,000	n.a.	×			×	
Breakfast Shopper	Manufacturers' Marketing Services, Inc.	16,500	197,958	×			×	
Know-It-All	Bristol-Myers Co.	11,084	n.a.	×			×	
Consumer Poll	Lever Brothers Co. & Miles Laboratories, Inc.	7,000	n.a.			×		×
Superstar Vacation	Westinghouse Electric Corp.	3,675	182,195		×			×
TOTALS				6	5	4	9	6

[a] Provided or estimated: bonuses were excluded when special coupons were required to be eligible for them.

[b] Total value was calculated from the maturity value of the bonds offered. It should be noted that in these two contests, total value excluded additional prizes that were awarded to youth organizations since the entrants did not directly benefit from them.

TABLE 2—SWEEPSTAKES RANKED BY SINGLE ENTRY GAIN EXPECTED UNDER MULTIPLE-PRIZE AWARDS PERMITTED PER CONTESTANT CONDITIONS

Sweepstakes	Expected Winnings From Single Entry	E_y to be Submitted to Maximize Expected Profit ($C=\$0.20$[a])	Expected Total Revenue	Total Cost	Expected Profit[b]	Expected Return[c]
Lysol Products	$0.30000	33,712	$8,257.71	$6,742.40	$1,515.31	22.5
The Great Meow Mix Meow Off	0.30000	112,372	27,525.43	22,474.40	5,051.03	22.5
G.E. New Car & Bike	0.20686	5,949	1,210.03	1,189.80	20.23	1.7
Miss America	0.19875					
Help Young America Save Energy	0.17857					
Glass Plus	0.10337					
Miss Clairol Silver Anniversary	0.10023					
Feetstakes	0.09195					
Breakfast Shopper	0.08335					
Help Young America	0.02544					
Superstar Vacation	0.02017					

[a] Calculated from equation (3).

[b] Calculated from equation (2).

[c] Expected Profit/Total Cost (shown in percent).

TABLE 3—SWEEPSTAKES RANKED BY SINGLE ENTRY GAIN EXPECTED UNDER ONE PRIZE PER ENTRANT AND TOP PRIZES DRAWN FIRST CONDITIONS

Sweepstakes	Expected Winnings from Single Entry	E_y to be Submitted to Maximize Expected Profit ($C=\$0.20$)	Expected Total Revenue	Total Cost	Expected Profit	Expected Return[a]
Lysol Products	.30000	208	$50.66	$41.60	$9.06	21.8
The Great Meow Mix Meow Off	.29814	1,004	237.32	200.80	36.52	18.2
G.E. New Car & Bike	.20686	37	7.53	7.40	.13	1.8
Miss America	.19875					
Help Young America Save Energy	.17857					
Glass Plus	.10337					
Miss Clairol Silver Anniversary	.10023					
Feetstakes	.09195					
Breakfast Shopper	.08335					
Help Young America	.02544					
Superstar Vacation	.02017					

[a] Expected Profit/Total Cost (shown in percent).

one's time in terms of $2.65 per hour (the then-minimum wage) and direct out-of-pocket costs per entry, we arrive at the not-unreasonable figure of 20¢ per entry.[4] Using equation (3) the optimal value of E_y is then presented in the third column of Table 2 for those contests satisfying entry conditions for a risk-neutral participant, that is, where expected profit is nonnegative for $E_y=1$. The corresponding values of E_y in Table 3 are obtained by optimizing π, equation (6) less CE_y, by means of a Fibonacci search procedure. Moreover, as expected, if one can win *all* prizes, optimizing behavior implies submitting many more entries then if one were eligible for only a single prize (compare the third columns of Table 2 and Table 3). What is a little surprising however, is the fact that, for both methods of prize eligibility, the expected return (the ratio of expected profit to total cost) is virtually the same (compare the corresponding columns in Tables 2 and 3).[5]

These conclusions must be qualified by the fact that they are based on the model given by equation (4). This assumes, we recall, that each entrant (other than the user of the model) submits only one entry (note the quantity $(i-1)$ in both numerator and denominator of (5)). While other entry numbers can be assumed (including the simplification of an average number of entries per every "other" entrant), the reader can see by inspection of (4) that burdensome complications can occur if other entry numbers are not equal. Another qualification is that our estimate of E_0 is based on perfect foresight, a quality that the typical contestant does not have. For our purposes, however, neither the single entry restriction nor the *ex post* values of E_0 are unreasonable.

Assuming that the average entrant can estimate E_0 with even ball park accuracy, the evidence is overwhelming that the bulk of contestants are either risk seekers, or participate for nonpecuniary motives, or

[4] Where multiple entries are permitted, this 20¢ estimate for the first entry is obviously too low. Frequently the cost of the first entry is substantially in excess of the unit cost of each subsequent entry. The opportunity cost of time devoted to absorb rules, restrictions, and constraints, to evaluate prizes, and even to obtain rough estimates of success probabilities can be considerable. In addition, sweepstakes rules generally require a separate mailing for each entry, thus entailing a minimum unit cost of postage, stationary, and labor. Consequently, while a cost of 20¢ for each entry after the first is in order, first-entry costs typically exceed those of subsequent entries and our results must be so qualified.

[5] Since all calculations in Tables 2 and 3 assumed a unit cost of 20¢ for each entry (including the first), total costs incurred to optimize E_y as well as the optimal value of E_y are overstated slightly.

both. As shown by Tables 2 and 3, less than one-third of the fifteen contests could attract either risk-neutral or risk-averse entrants. All other sweepstakes would be avoided by such people. Yet these contests drew 3,610,000 entries out of 4,610,000, the total number of entries for all surveyed contests. Even allowing for gross errors in estimating E_0 or the value of some prizes, one cannot help but be impressed by the inconsistency of this evidence against the hypothesis that most entrants are either risk neutral or risk averse.[6]

As is well known, by appropriately expanding the utility function we can rationalize the entry of nonrisk seekers in unfair bet contests, including both sweepstakes and lotteries. To be precise, we specify an individual utility function that depends not only on the individual's wealth but also on the activities of gambling and "just doing something," that is, occupation of idle time. The first partial derivatives of utility with respect to each of these variables is assumed to be positive even though we have reservations about this property holding for moderately large positive values of the two activities, for example, it seems unreasonable to assume the typical individual's pleasure from gambling is unbounded or that an inordinate amount of time would be devoted to such activities. In sum, the axiom of nonsatiation does not extend to these activities.

It is plausible to assume that among certain people the response of utility to these two activities will differ substantially. There is reason to believe that regular sweepstakes' contestants may be motivated more by the desire to occupy idle time than to obtain pleasure from gambling. An entry in a sweepstakes contest is often time consuming. Unlike sweepstakes participation, purchase of a pure lottery[7] ticket requires typically a very small amount of time (a possible exception is when a participant must, in order to buy a ticket, engage in travel which otherwise would not be made). Therefore, people who enter sweepstakes are more likely to do so to occupy idle time than those who purchase lottery tickets. If so, then a significant fraction of sweepstakes entrants do not or would not participate in pure lotteries.[8]

III. Summary

Information provided by sweepstakes advertisements is severely limited. None of those surveyed hinted as to their expected number of entries although one did state the estimated odds of winning. (*Ex post* that estimate was in substantial error, doubtless leaving the sponsor vulnerable to potential investigation.) Moreover, the value of individual and total prizes was frequently difficult, if not impossible, to discern. Finally, not one of the advertisements indicated the order in which the prizes were to be drawn. Thus, in virtually every contest it was not possible to establish precise entry conditions for the rational entrant. Inferences had to be drawn as to the likely value of many prizes, estimates were required of E_0, and guesses made of the order in which prizes were to be drawn.

We hope the above will not entice the already overburdened Federal Trade Commission into action, but sweepstakes sponsors could enhance the attractiveness of their contests by supplying more details on prize values, more information on the drawing procedure, and at least, the number of entries in prior similar contests. To maxi-

[6] Another "noise" contributing factor is the evaluation we were forced to make of prizes that were expressed in kind. Contestants would be expected to place different monetary valuations on such prizes and, indeed, even omit the operation by implicitly expressing utility functions in terms of the attributes of the prizes.

[7] A pure lottery is defined as one in which all net proceeds (i.e., receipts less prizes) are retained by the sponsor and no distribution is made to select groups, for example, the indigent or the elderly.

[8] A definite test of this hypothesis remains to be undertaken. However, a preliminary sample of seven sweepstakes winners (which happens to be at hand) does not lend support to this hypothesis—all seven indicated either that they would participate in state-run lotteries if such were available (note that such lotteries operate only in certain states), or that they currently participate. To be sure, some state-run lotteries distribute portions of their proceeds to specific demographic groups such as the elderly. Thus there may be a philanthropic motive as well among these players.

mize both the number of individual entries and the number of entries per participant in single prize contests, the sponsor should maximize the ratio S/C. Directing efforts to reduce C may be much more effective than increasing S. Some acts of reducing C could have a direct cost to the sponsor such as free postage for a contestant's first entry; others, however, might be costless such as providing more information on prizes, on the order of drawings, and on the number of entries attracted to prior contests.

Of the contests that could be analyzed, over two-thirds could not attract so-called risk-neutral or risk-averse entrants. As an alternative to the implication that most participants are risk seekers, attention was called to the well-known belief that these people may be dominated by nonpecuniary motives: pleasures of gambling or use of leisure time. Although little has been done by economists to test the strength of these motives in individual decision making, it was plausibly suggested that a large fraction (significantly different from zero) of sweepstakes players would not participate in pure lotteries. Evidence from a small sample did not support this hypothesis, but it is likely that it was contaminated with additional player motives.

Finally, it should be recognized that sweepstakes contests are conducted for the purpose of advertising a firm's image, its name, or some of its specific products. Therefore, advertising of a sweepstakes is literally advertising about advertising, which in a sense, is a kind of meta ,advertising. Since no product or service is literally being offered for sale—merely a no-explicit-entry-fee contest—regulation of meta advertising may be beyond the authority of regulatory agencies.

REFERENCES

Donald P. Davidson, Patrick Suppes, and S. Siegel, *Decision-Making: An Experimental Approach*, Stanford 1957.

M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," in *Readings in Price Theory*, Chicago 1952.

F. C. Mosteller and P. Nogee, "An Experimental Measurement of Utility," *J. Polit. Econ.*, Oct. 1951, *59*, 371–404.

A Monopoly Model of Public Goods Provision: The Uniform Pricing Case

By Geoffrey Brennan and Cliff Walsh*

This paper offers a contribution to the small literature that has emerged over the last decade on what are sometimes called marketable; or price-excludable, public goods, but which for reasons of convenience and economy we shall henceforth term "joint goods."[1] These goods combine the "Samuelsonian jointness of supply" (or nonrivalness in consumption) characteristic of a pure public good with the "costless excludability" characteristic of a pure private good: all units produced could be consumed fully and equally by all, but individuals can be excluded from consuming any or all units for which they choose not to offer to pay.

Interest in such goods stems from two sources. First, joint goods *are* frequently provided by the market. For example, many transportation, recreation, and entertainment services exhibit substantial jointness elements but direct pricing can be, and often is, practiced, while copyrights and patents perform the function of exclusion devices in relation to the outputs of writers and researchers. The important exercise of extending traditional micro-economic analysis to incorporate joint goods, however, remains substantially incomplete, especially in the context of monopoly provision.

Second, the question naturally arises of whether the efficiency characteristics of market outcomes for joint goods more nearly correspond to those of public goods or to those of private goods. Since Samuelsonian jointness implies that any individual can be allowed to consume units of a joint good without cost to existing users, optimal market provision involves the demanding requirement that all individuals should consume total output (an outcome that *necessarily* prevails for public goods) with that output determined by the familiar public good rule, "$\Sigma MRS = MRT$." On the other hand, the possibility of (costless) exclusion in the joint good case eliminates the financing problem faced by producers of pure public goods, and there is therefore no *necessary* presumption that market failure will be as marked for joint goods as for public goods. Examination of this issue may, indeed, throw light on the question of whether Samuelsonian jointness or nonexcludability constitutes the more virulent source of market failure and thus may contribute to a more general normative theory of public sector intervention.

Models of competitive provision dominate the existing literature. Earl Thompson (1968) assumes that all buyers and sellers have perfect information on all buyers' preferences and consumption levels, and observes that in this context price discrimination is possible even with competitive provision of joint goods since there is no competition between consumers over any particular unit of output. Hence profit-maximizing firms will sell access to all units they produce at the highest per unit price each consumer is prepared to pay, but freedom of entry results in profits being driven to zero through a significant *overexpansion of output*.

*Professor of economics, Center for the Study of Public Choice, Virginia Polytechnic Institute and State University, and professor of economics, University of Adelaide, Australia, respectively. The paper was originally prepared while we held positions at Australian National University, and earlier versions, widely circulated in various forms (see our 1978 paper), particularly benefited from discussions with John Head and Ted Seiper, both of whom can claim joint property rights in some of the results. The final version has been improved by the efforts of an anonymous referee and George Borts. Nonetheless we accept the conventional attribution of "liability for errors."

[1] The contributions of Earl Thompson (1968) and William Oakland are by far the most significant in the published literature. However, see also the work of Harold Demsetz, Richard Auster, Dwight Lee, and our 1979 paper, plus Thompson's (1969) response to the totally misplaced comments on his original piece. Of all these papers, only those by Lee and ourselves are explicitly concerned with monopolistic behavior.

William Oakland, on the other hand, adopts the more usual assumption that sellers have no more information than the market process normally generates for them and observes that, with buyers and sellers responding parametrically to price, zero profit on any unit must involve a price of MC/m, where m is the number of consumers of that particular unit. Allowing m to vary from n to 1 (where n is population size) generates a step function of prices faced by all consumers in which unit prices vary inversely with intensity of use. Since under this price arrangement the marginal unit will not in general be consumed by all, *output will be suboptimal and consumption more so.*[2]

Discussion of monopoly models, however, has been surprisingly slight in view of the variety of alternative pricing strategies that might be considered relevant. Thompson, in the course of his analysis, briefly considered the logical counterpart to his competitive model—the case of perfectly discriminating monopoly (a case also discussed, less formally, by James Buchanan). The firm again appropriates full consumer's surplus as profit, but with profit now maximized (rather than driven to zero by competition), total output and the consumption level of each individual is Pareto optimal. Thus, in the Thompson world, monopoly performs much better than competition in an efficiency sense. The appropriate monopoly analogue to the Oakland model is, however, much less obvious since even with limited information monopoly firms may choose a variety of pricing strategies. In our 1979 paper, we force a precise analogue on the monopoly producer: that is, he is constrained to seek maximum profit while setting prices which vary inversely with intensity of use. While the resulting monopoly outcome is necessarily less efficient than the Oakland competitive outcome, the solution strains the Oakland assumptions about information because at least for marginal units (the price of which in turn affects prices for inframarginal units) the seller is dealing with small numbers of buyers (possibly only one) who cannot be expected to behave atomistically.[3]

In this paper we direct attention to a model which is both an alternative monopoly counterpart to Oakland's competitive model and the most logical extension of the familiar private good monopoly model. That is, we suppose that in response to limited information about consumer preferences (or because the firm is subject to regulation which prohibits discrimination of any form), the monopoly producer of a joint good chooses to set a price which is *uniform* both over quantity and between individuals. The formal analysis is set out (in Section I) through a series of propositions which indicate the more striking features of the solution and identify some simple (though very unusual) comparative static results and welfare properties. Section II illustrates the main results in a diagrammatic example, while Section III offers a concluding summary. The particular framework of analysis we have adopted is designed to highlight the similarities, but especially the differences, between private good and public good monopoly provision.

I. The Formal Analysis

A profit-maximizing monopolist offers for sale, at a uniform per unit price, a joint good G, defined such that while any individual can be costlessly excluded from consumption of all or any units of G produced, once a unit is provided to any one individual, it can be costlessly provided to all others in the relevant group without reducing consumption by any. Thus

$$(1)\qquad G_i \leqslant \bar{q} \qquad \text{for all } i = 1, \dots, n$$

where G_i is the ith individual's actual con-

[2] Demsetz and Auster also present competitive models both of which, in different ways, claim to show that full optimality will emerge from competitive provision. However, John Head has effectively disposed of the Demsetz claims, and the Auster results appear equally susceptible to fundamental criticism.

[3] Lee's analysis of all-or-nothing pricing (but with identical all-or-nothing offers to all consumers) though incomplete provides a highly suggestive alternative basis for a monopoly model with limited information. For a further development of this case, see our paper with Lee.

sumption of G, $\bar{q}$ is production of G, and n is population size.

For ease of exposition, we order individuals in terms of intensity of demand for G, so that[4]

$$(2)\quad q_1(p) \leqslant q_2(p) \leqslant \ldots \leqslant q_n(p) \quad \text{for all } p$$

Moreover, except where otherwise stated, we assume that average (and hence marginal) production costs are constant, and to facilitate welfare comparisons we assume all other goods are purely private and produced under competitive conditions.

Our objective is to isolate the conditions for a profit-maximizing solution and to indicate some of the more striking properties of that solution. Given the relatively unfamiliar aspects of the model's structure, we proceed by establishing a series of basic propositions. As an important point of departure, however, we should note that (in contrast to the familiar private goods case where in equilibrium units in production and units in consumption are identical) in the joint goods context, total consumption is the aggregate of individual consumptions but total production is the maximum of them:

$$(3a)\qquad \bar{q} = \max_i G_i$$

while any individual's consumption of G is given by the relation

$$(3b)\qquad G_i = \min\{q_i(p), \bar{q}\}$$

where G_i is actual consumption, $q_i(p)$ is quantity demanded at the uniform per unit price p, and $\bar{q}$ is output. Since costs are a function of output while revenue is a function of price and aggregate consumption, the simple $MC = MR$ rule—while it remains conceptually valid—is bound to involve complications not present in the private goods case. This observation suggests an initial proposition.

PROPOSITION 1: *Average revenue exceeds price for a joint good.*

DISCUSSION:

In the private good case, where aggregate consumption and production are equal, average revenue and price are identical under uniform pricing; i.e.,

$$R = p\sum_{i=1}^{n} x_i = p\bar{q} \quad \text{and hence } \frac{R}{\bar{q}} = p$$

However, in the joint goods case,

$$(4)\quad \sum_{i=1}^{n} G_i > \bar{q} \quad \text{and} \quad \frac{R}{\bar{q}} = \frac{p\Sigma G_i}{\bar{q}} > p$$

This occurs because in the joint good profit-maximizing equilibrium some individuals (possibly only one) will be consuming total output, while the remainder will be allowed access to however much they demand at the ruling uniform price. Average revenue must exceed price.

PROPOSITION 2: *It may be, and in general for some levels of marginal cost will be, profit maximizing for the firm to ration output—that is, to choose an output/price combination which leaves some individuals consuming less than they demand at the ruling price.*

DISCUSSION:

Suppose, contrary to the proposition, that rationing occurs by price alone. Then output is given by $\bar{q} = q_n(p)$, and profit by

$$(5)\quad \Pi = R - C = \sum_{i=1}^{n} pq_i(p) - C[q_n(p)]$$

Differentiating (5) with respect to p yields

$$(6)\quad \frac{d\Pi}{dp} = \sum_{i=1}^{n} \frac{d(pq_i)}{dq_i} \cdot \frac{dq_i}{dp} - \frac{dC}{dq_n} \cdot \frac{dq_n}{dp}$$

Or, denoting dq_i/dp (the slope of i's demand curve) by S_i, interpreting MR_i as the ith individual's marginal revenue, and setting (6) at zero, the conditions for a profit maximum can be stated as

$$(7)\qquad \sum_{i=1}^{n} \frac{MR_i S_i}{S_n} = MC$$

[4] For this ordering to be invariant with respect to price, it is necessary either to assume that demand curves do not intersect, or to allow the identity of the consumers to change with changes in p. There seems no reason not to permit the latter, but for expositional reasons we have assumed the former.

The left-hand side of (7) defines the relevant expression for market marginal revenue in output units, and differs from the private goods case in that S_n enters the denominator (rather than ΣS_i) because each production unit can be consumed by all.[5]

The assumption of rationing by price alone ensures that for any prospective output level $\bar{q}$, price is given by $p=p_n(\bar{q})$ determined from the (inverse) demand function of n. However, for some levels of $\bar{q}$ it is possible that

$$(8) \qquad \sum_{i=1}^{n} \frac{MR_i S_i}{S_n} > p_n(\bar{q})$$

That is, the market marginal revenue curve may lie above n's demand curve, a possibility which is made transparent by the observation that n's marginal revenue $MR_n(\bar{q})$ enters the left-hand side of (8) with a weight of one, so that (8) only requires that

$$(8') \qquad \sum_{i=1}^{n-1} \frac{MR_i S_i}{S_n} > p_n(\bar{q}) - MR_n(\bar{q})$$

and in general this becomes increasingly likely as $\bar{q}$ is decreased. But if (7) is satisfied at a level of $\bar{q}$ for which (8) is true then marginal cost will exceed per unit price.

For joint goods (unlike private goods), $p<MC$ is not only possible but may be profitable since in general some units are consumed by many individuals. But if $q_n(p) > q_i(p)$ (for $i \neq n$), the nth individual alone consumes the marginal production units, and if $p<MC$, the profit obtained from those marginal units is negative: profit could be increased by reducing output without any change in price. Hence, contrary to the supposition adopted to generate (7), it cannot be profit maximizing always to supply all individuals with the quantities they desire at the ruling price. Rationing by output may increase profit, as stated.[6]

Given this, a more general formulation of the producer's problem is suggested. We can now write revenue as

$$(9) \qquad R = \sum_{i=1}^{k} pq_i(p) + (n-k)p\bar{q}$$

where $(n-k)$ is the number of individuals rationed by output, and profit is

$$(10) \quad \Pi = \sum_{i}^{k} pq_i(p) + (n-k)p\bar{q} - C(\bar{q})$$

As the earlier discussion makes clear, a necessary condition for profit maximization is that

$$(11) \qquad (n-k)p \geqslant MC$$

For given MC, this represents a restriction on both price (p) and the number of rationed individuals $(n-k)$. The profit-maximizing conditions must therefore involve MC, p, and k, with the latter two effectively determining output $\bar{q}$. In this connection we may now consider a further proposition.

PROPOSITION 3: *In the constant costs case, the locus of potential equilibrium combinations of price and output consists of a set of disjoint portions of individual demand curves.*

DISCUSSION:

The proof involves the observation that any profit-maximizing price/output combination must lie on *some* individual's demand

[5] In private goods monopoly analysis, we usually take the convenient approach of working with aggregate output (= aggregate consumption) and suppressing the summations explicitly identified here. Were we to do otherwise, the profit function for private goods would be as in (5) except for the fact that $C = C[\Sigma q_i(p)]$, which results in ΣS_i rather than S_n appearing in the market marginal revenue function. Alternatively, note that in the joint good case, (6) could have been rewritten in terms of consumption units as $\Sigma MR_i S_i / \Sigma S_i = S_n MC / \Sigma S_i$: the left-hand side is identical to the private good expression for market marginal revenue, while the right-hand side signifies the fact that each production unit (with cost MC) generates many consumption units. Note that $MR_i = MR_i[q_i(p)]$ and $S_i = S_i[q_i(p)]$: i.e., they are evaluated at the different quantities consumed by different individuals at a given uniform price.

[6] The rationing outcome and the analytical circumstances in which it arises were discussed in more detail in an earlier version of this paper, and are illustrated for one particular case in Section II below. The rationing result has now been discussed in alternative analytical frameworks by Michael Burns and Walsh, and by Dagobert Brito and Oakland.

curve. For, suppose the contrary. We can differentiate (9) with respect to output to obtain

$$\frac{dR}{d\bar{q}}=\sum_{i=1}^{k}\frac{dR_i}{dp}\cdot\frac{dp}{d\bar{q}}+(n-k)p+(n-k)\bar{q}\frac{dp}{d\bar{q}} \tag{12}$$

Now, a necessary though not sufficient condition for profit maximization is that the price set at any output level $\bar{q}$ maximizes revenue. Hence, we can differentiate (9) with respect to price to obtain

$$\left.\frac{dR}{dp}\right|_{q=\bar{q}}=\sum_{i=1}^{k}\frac{dR_i}{dp}+(n-k)\bar{q}=0 \tag{13}$$

which with (12) gives us

$$\left.\frac{dR}{d\bar{q}}\right|_{p=p^*}=(n-k)p^* \tag{14}$$

And profit maximization over the range in which $\bar{q}\neq q_k$ for any k requires that $(n-k)p^*$ is equal to marginal cost.

This result can be justified heuristically as follows. Suppose that the profit-maximizing price is p^*, and the number of individuals rationed is $(n-k)$, so that $q_k(p^*)<\bar{q}<q_{k+1}(p^*)$. Then: (a) $(n-k)p^*>MC$, or (b) $(n-k)p^*<MC$, or (c) $(n-k)p^*=MC$. Clearly, in cases (a) and (b) output adjustment is profitable: in the former, adding output up to $q_{k+1}(p^*)$ adds $[q_{k+1}(p^*)-\bar{q}]$ $[(n-k)p^*-MC]$ to profit, while in the latter, cutting output to $q_k(p^*)$ adds $[\bar{q}-q_k(p^*)]\,[MC-(n-k)p^*]$ to profit. Only case (c) is a *potential* profit maximum.

Now we can show that $MR_{\bar{q}}$ is *increasing* over the range in which $\bar{q}(p^*)\neq q_k(p^*)$ for some k, by showing that as $\bar{q}$ rises in this range so must p^*. Equation (13) requires that

$$(n-k)\bar{q}=-\sum_{i=1}^{k}\frac{dR_i}{dp} \tag{15}$$

so that if $\bar{q}$ increases, so must the absolute value of $\Sigma(dR_i/dp)$ (which must be negative). Given conventional demand conditions, this expression increases with increased price, being simply the aggregate marginal revenue with respect to price for the first k individuals. Hence

$$\frac{dp^*}{d\bar{q}}>0 \tag{16}$$

and, from (14) above

$$\frac{dMR_{\bar{q}}}{d\bar{q}}=(n-k)\frac{dp^*}{d\bar{q}}>0 \tag{17}$$

But if marginal revenue is *increasing* over the range in question, and given constant marginal and average costs, then no equality between marginal revenue and marginal cost can be profit maximizing in that range: a small increase in output leads to a small increase in revenue-maximizing price and, with no change in marginal cost, a corresponding increase in profit. Only if marginal cost is increasing and increasing faster than marginal revenue can the relevant second-order conditions be satisfied.

The final observation necessary to establish our proposition is that, by assumption, the demand curves of the n individuals do not intersect. Hence, the locus of potential equilibria—the portions of different individual's demand curves—must be discontinuous in the constant cost case, and *a fortiori* for decreasing costs.

PROPOSITION 4: *As marginal costs rise, price may fall over some ranges (specifically, the ranges in which it becomes profitable to have output determined by a lower demand curve—i.e., ration additional consumers).*

DISCUSSION:

Suppose we are on individual k's demand curve. Then profit can be written:

$$\Pi=\sum_{i=1}^{k}pq_i+(n-k)pq_k-C(q_k) \tag{18}$$

which yields, as the profit-maximizing condition,

$$\sum_{i=1}^{k}MR_iS_i+(n-k)MR_kS_k=S_kMC \tag{19}$$

using the same notation as in equation (7) above. Equation (19) defines the profit-maximizing price, p_k^*: output is then determined from k's demand curve. The obvious explanation for (19) in relation to (7) is that all the "rationed" individuals are treated *as if* they had exactly the same demand for G as the kth individual.

Now a necessary condition for the profit-maximizing solution to lie on k's demand curve is that

$$\text{(20a)} \qquad (n-k+1)p_k^* > MC$$

and

$$\text{(20b)} \qquad (n-k)p_k^* < MC$$

If (20a) does not hold, then we should be rationing more consumers; and if (20b) does not hold, we should be rationing fewer.

Likewise, in the range in which we lie on the $(k+1)$th individual's demand curve,

$$\text{(21a)} \qquad (n-k)p_{k+1}^* > MC$$

and

$$\text{(21b)} \qquad (n-k-1)p_{k+1}^* < MC$$

At the point at which it just becomes profitable to ration the $(k+1)$th individual, it must be true that, at that level of marginal cost, profit is the same whether we are producing q_{k+1}^* at p_{k+1}^*, or producing the smaller quantity q_k^* at the price p_k^*. Let the level of marginal cost equal $\overline{MC}$. Then,

$$\text{(22)} \quad (n-k)p_k^* < \overline{MC} < (n-k)p_{k+1}^*$$

from (20b) and (21a), from which it follows that $p_{k+1}^* > p_k^*$. Thus, just below $\overline{MC}$, it is profitable to produce on $(k+1)$'s demand curve; just above $\overline{MC}$, it is profitable to produce on k's demand curve. And in the neighborhood of $\overline{MC}$, the profit-maximizing price will *fall* as MC rises.[7]

It should be noted that what we have established here is that, for "small" increases in marginal cost, in the vicinity of the changeover point to more severe rationing, price may fall. Overall, however, the price-output schedule must be downward sloping so that "large" increases in marginal cost will be associated with price increases, though these price increases too will *ceteris paribus* be less than with private goods. How large is large and how small is small depends on the characteristics of the distribution of demands for the joint good. However, it has been shown elsewhere that even with a continuous distribution of demand curves, the price-output schedule may be positively sloped over some ranges (but not overall).[8] In any event, even if it applies only over limited ranges, this result is striking and unusual, and reflects the fact that models of joint good provision have quite unfamiliar structure.

PROPOSITION 5: *The uniform price monopoly solution is Pareto inferior to the Oakland competitive solution.*

DISCUSSION:

Suppose, without loss of generality, that marginal costs are such that it is profit maximizing to choose a price-output combination on the kth individual's demand curve. Then, from (20a) above,

$$p_k^* \geqslant \frac{1}{n-k+1} MC$$

In the Oakland competitive solution,

$$\text{(23)} \qquad p_j^c = \frac{1}{n-j+1} MC \quad \text{for all } j$$

where p_j^c is the price charged on the bundle of units with $(n-j+1)$ consumers in the Oakland solution. Therefore,

$$\text{(24)} \qquad p_j^c < p_k^* \quad j=1,\dots,(k-1)$$

$$\text{(25)} \qquad p_k^c \leqslant p_k^*$$

[7] This proof depends on constant costs. It would be valid a fortiori under increasing costs, but would not hold if marginal costs are decreasing sufficiently rapidly over the relevant output range.

[8] The interested reader is referred to Burns and Walsh for a discussion using an alternative, demand distribution, approach.

Hence, individuals $1,\ldots,(k-1)$ consume more G at lower prices, and k consumes at least as much at a (marginal) price no higher, under the Oakland competitive solution. Consider now individuals $(k+1),\ldots,n$. If p_k^* equals p_k^c *and* p_k^c is the highest price in the Oakland price function, they consume no less at a (marginal) price no higher under the Oakland solution (though they clearly gain from lower prices on infra-marginal units). However, if $p_k^c<p_k^*$ and/or p_k^c is not the highest Oakland price, then they will consume more, though for some of the *additional* units they may pay prices higher than p_k^*. Nonetheless, in this latter case since they are free to choose not to consume the additional units, any additional purchases must be welfare improving, too.

Thus, consumption of G (freely chosen) under the Oakland competitive solution is, for each and every consumer, greater than or equal to his consumption under the uniform price monopoly solution. Since there is, in general, underconsumption in the Oakland model, there is greater underconsumption with monopoly, and the Oakland solution is Pareto superior to the uniform price monopoly solution.

II. An Illustrative Example

The salient features of the joint goods monopoly model can be illustrated in a simple diagrammatic example, involving a two-consumer economy. In Figure 1, D_1 and D_2 represent the demand curves of individuals 1 and 2 for a joint good G. For simplicity, both demand curves are drawn linear and have the same intercept along the vertical axis to avoid problems of discontinuity in the aggregate marginal revenue schedule. Individual 1 demands one-third the quantity of G that individual 2 demands at any price. On this basis, the average revenue curve, if there is no rationing, is depicted by AR_N. At any price, output will be the quantity demanded by individual 2, and total revenue is $R=pq_2+pq_1$. Since (by construction) $q_1=(1/3)q_2$ and (without rationing) $\bar{q}=q_2$:

$$R=\frac{4}{3}pq_2 \quad\text{and}\quad \frac{R}{\bar{q}}=\frac{4}{3}p(>p)$$

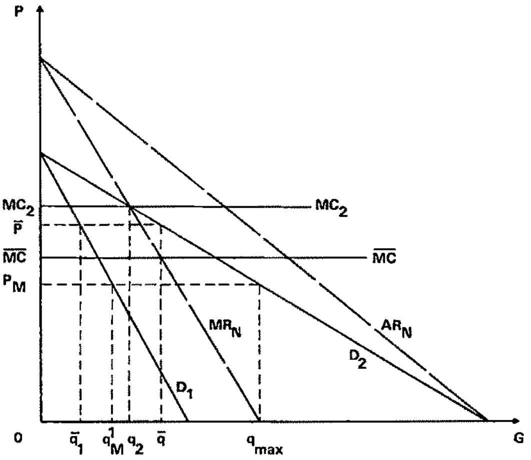


FIGURE 1

Given the linearity of the underlying demand curves, AR_N is also linear. We can then derive from it a marginal revenue curve MR_N, which is in fact a graphical depiction of the left-hand side of (7) above.

Now we can isolate the range in which MR_N and D_2 are potentially relevant—that is, the range over which rationing does not necessarily occur. First, it can never be profit maximizing to produce more than q_{max}, where MR_N is zero: this is the output at which total revenue is maximized. At this point, price is p_M derived from D_2 at that output, and individual 1 consumes q_M^1. As marginal cost rises from zero, output is determined where marginal cost equals MR_N, and price is derived from D_2. Thus, where marginal cost equals $\overline{MC}$, output will be $\bar{q}$, price $\bar{p}$; individual 2 will consume $\bar{q}$ and individual 1, $\bar{q}_1$. Second, however, at MC_2, MR_N and D_2 intersect. As marginal costs rise above this level, the price generated from D_2 will lie *below* marginal cost: *rationing* is clearly now required (though later discussion will show this point is not always the critical changeover point). Thus, the no-rationing regime is potentially relevant over the range q_{max} (zero marginal cost) to q_2 (with MC_2): over this range, output is determined where MC and MR_N intersect, and price determined from D_2.

With only two consumers, there is only one "rationing" outcome: that in which in-

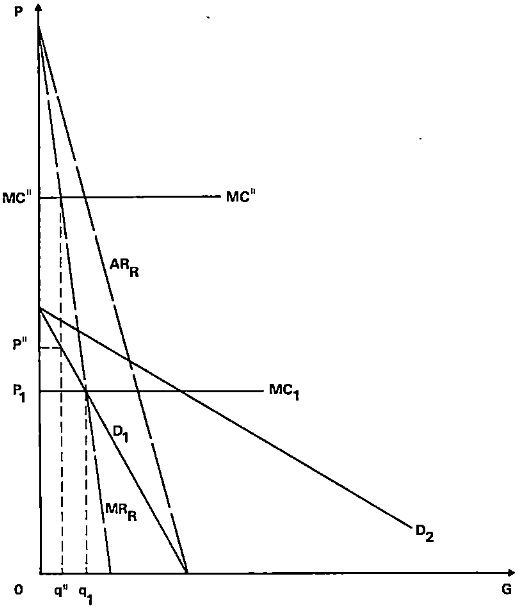


FIGURE 2

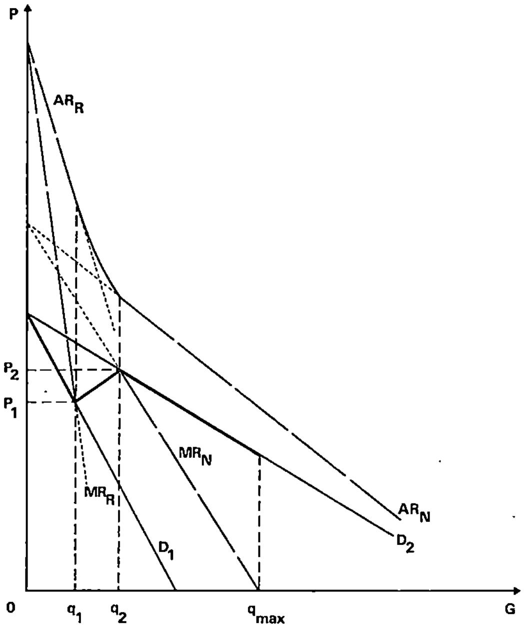


FIGURE 3

dividual 2 is rationed. For a start, let us focus on the case where output and price are determined along individual 1's demand curve: that is, individual 2 is being treated *as if* he had individual 1's demand curve and average revenue is twice D_1 in a vertical direction.

$$R=2pq_1(p)=2p\bar{q} \quad \text{and} \quad \frac{R}{\bar{q}}=2p$$

This is shown as AR_R in Figure 2; and the corresponding marginal revenue curve MR_R is derived from AR_R (which is linear) in the standard convenient way. As we have seen, a necessary condition for being on D_1 is that per unit price is less than marginal cost: if this is not so, then rationing cannot be profitable. Thus, the range of outputs for which D_1 is potentially relevant is the range from zero to q_1. At q_1, MR_R and price are equal, so that a marginal cost equal to MC_1 would yield price p_1 and output q_1. For all levels of marginal cost above MC_1, the price determined along D_1 lies below marginal cost. For example, at MC'' output is q'' where MC'' and MR_R intersect; price is p'', less than MC''; and both 1 and 2 consume q''.

We have in this simple diagrammatic example isolated a range in which D_2 is potentially relevant—from q_{max} to q_2—and a range in which D_1 is potentially relevant—from q_1 to zero. But what of the range between q_1 and q_2? This is the range in the neighborhood of the intersection of AR_R and AR_N (as shown in Figure 3). As equation (14) above shows, over this range, marginal revenue and price are identical,[9] and the marginal revenue schedule slopes upward throughout. Thus, the marginal revenue schedule takes the form of a line connecting (p_1, q_1) and (p_2, q_2) and in this case of linear demands, the relevant line is straight.[10] Average revenue over this same range follows neither AR_R nor AR_N: it is, rather, a continuous convex portion as shown in Figure 3, meeting AR_R at q_1 and

[9] Given that only one person is rationed. More generally, $MR=m\cdot p$, where m is the number of individuals "rationed."

[10] Over this range of output we have $R=p\cdot q_i(p)+p\bar{q}$, where $q_i(p)$ is of the form $q_i(p)=a-bp$. Substituting into R, and differentiating with respect to price yields $dR/dp=a-2p+\bar{q}$. Setting this at zero reveals p^* to be a linear function of $\bar{q}$ over the relevant range: i.e., $p^*=(\bar{q}+a)/2b$.

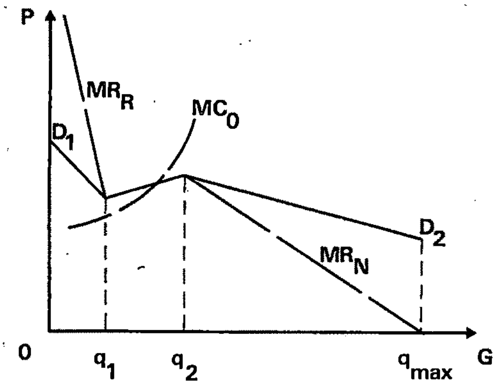


FIGURE 4a

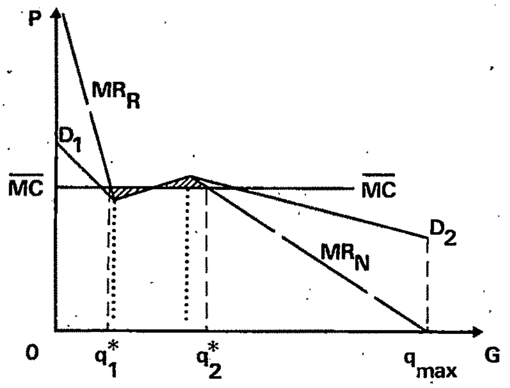


FIGURE 4b

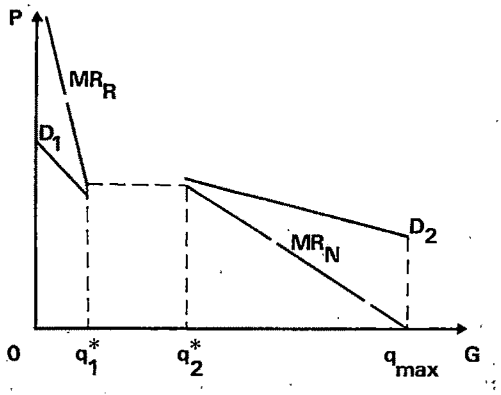


FIGURE 4c

AR_N at q_2, and lying above both over the q_1q_2 range.

The locus of marginal revenue curves and the corresponding portions of D_1 and D_2 plus the interconnecting linear portion over the range q_1 to q_2 represents, for this two-person example, a complete diagrammatic description of the potential equilibria. This is shown in Figure 4a. With the additional assumption of constant costs, we can discard some of these potential equilibria as irrelevant, for there exists a level of marginal costs at which rationing becomes profitable between p_2 and p_1. In fact, this is $\overline{MC}$ in Figure 4b at which profit is the same whether rationing is applied or not. At this level of marginal costs, the area under the marginal revenue schedule above the cost line is the same whether the MR_R segment or the MR_N segment is applied—that is, the two shaded areas in Figure 4b are identical. In this constant costs case, then, the set of potential equilibria is represented in Figure 4c, with D_2 applying over the range q_{max} to q_2^*, and D_1 applying over the range q_1^* to zero. As Proposition 4 indicates, an increase in marginal costs from just below $\overline{MC}$ to just above $\overline{MC}$ leads to a *reduction* in price —an outcome which, to our knowledge, occurs nowhere else in economics.

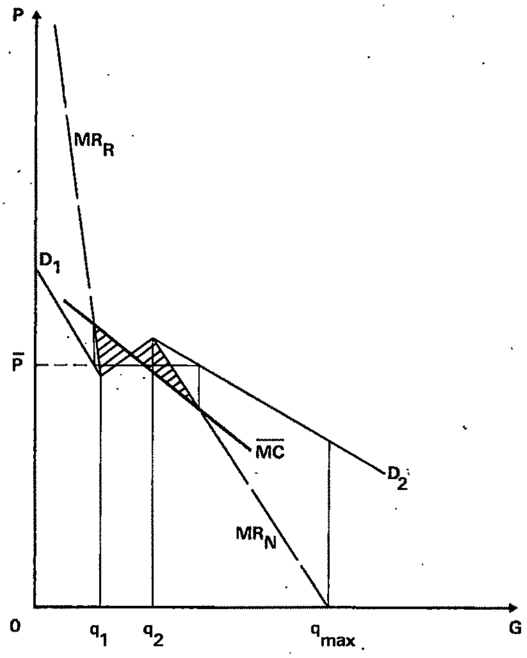


FIGURE 5

With more general cost conditions allowable, no part of the locus illustrated in Figure 4a can be discarded since a stable equilibrium along the upward-sloping portion is feasible if marginal costs are increasing sufficiently rapidly (as the MC_0 schedule is). Obviously, an increase in marginal costs can, in this case too, lead to a reduction in price: that result does not depend on the constant costs assumption. However, it does depend on marginal costs not *decreasing* at a sufficiently rapid rate. As illustrated through Figure 5, if marginal production costs are decreasing faster than indicated by the $\overline{MC}$ schedule, then price will rise as we move from D_2 to D_1: with $\overline{MC}$, profits are identical under the rationing and no-rationing regimes (i.e., the shaded areas are equal) and prices from D_1 and D_2 are also equal ($\bar{p}$). This in no way undermines the interest that our result can claim, of course: we can (and have) established its validity in this particular simple example for a variety of cost structures, and universally for constant costs.

III. Summary and Conclusions

The primary objective of this paper has been to develop a model of market provision of price-excludable public goods under conditions of monopoly, using standard assumptions about the information possessed by the firm. These assumptions are taken to preclude the possibility of any discrimination either between different consumers or over different units of the joint good, so that the monopolist charges a price which is identical between consumers and uniform over units.

The profit-maximizing equilibrium that we isolate turns out to have properties which contrast rather conspicuously with the analogous ones from private goods analysis. First, except where marginal production costs are "low" (including the zero cost case), the equilibrium will typically be characterized by rationing of the highest-demand individuals, in the sense that at the prevailing profit-maximizing uniform price those individuals demand more than it is profitable for the monopolist to supply to them. Second, whereas the locus of potential price-output equilibria for the private good case is the market demand curve, and under conventional assumptions is a continuous and monotonically decreasing function of price, an analogous locus in the joint goods case consists of a set of *discontinuous* portions of different *individual* demand curves, those of higher-demand individuals being relevant when marginal costs are low with those of successively lower-demand individuals becoming relevant as marginal costs increase. These portions of individual demand curves are connected by line segments that are necessarily upward sloping, and one important consequence of this is that, in the joint goods case, a rise in marginal costs may lead to a *reduction* in price. Heuristically, the reason for these results is that whereas in the private good case there is a one-to-one relationship between output and consumption, in the joint good case any particular level of output can be associated with a wide range of consumption levels, depending on the price charged. Thus, although total *consumption* and price behave in an entirely familiar fashion, *output* and price do not.

Thompson (1968) demonstrates that, under assumptions of "perfect information," monopoly is superior to perfect competition in the standard efficiency sense. The information assumptions used by us are more conventional, and are most akin to those adopted by Oakland in his treatment of the perfectly competitive case. Comparing our model with the Oakland model, it becomes clear that, as in the private good case, monopoly involves a *larger* welfare loss than does perfect competition. In this sense, contrary to Thompson's result, the possibility of monopoly provision in no way rescues the market from inefficiency in the provision of joint goods under conventional information assumptions. On the contrary, market failure is even more conspicuous with monopoly than with competitive market provision.

REFERENCES

R. D. Auster, "Private Markets in Public Goods (or Qualities)," *Quart. J. Econ.*, Aug. 1977, *91*, 419–30.

G. Brennan and C. Walsh, "A Monopoly Model of Public Goods Provision: The Uniform Pricing Case," seminar paper no. 70, Dept. Econ., Monash Univ., Jan. 1978.

Competitive Production and Increases in Risk

By STEVEN A. LIPPMAN AND JOHN J. MCCALL*

The theory of the firm is a monumental achievement of neoclassical economics. Without this "engine of analysis," it would be difficult, if not impossible, to comprehend the pricing, output, and input decisions of firms as they respond to routine events like the imposition of a tax, the opening of a new market, a technical innovation, and a sudden shortage of a key factor of production. It is remarkable that this theory has been successful in explaining behavior that to a large extent is motivated by both profit and risk when the theory itself has only explicitly considered the profit motive. This is not the place to dwell on the evolution of economic theory; it suffices to note that there are many important economic phenomena that the purely deterministic theory does not explain.[1] The purpose of this note is to elucidate the way in which risk influences the output decisions of risk-averse entrepreneurs and the number of risk-neutral firms in a competitive industry. In both cases, the predicted behavior differs from that of the deterministic theory.

In our study we shall restrict attention to the behavior of firms in a single period setting.[2] The firm has no control over price and, because storage makes no sense, simply sells all of its output at the going price. The source of uncertainty is the requirement that the firm produce before price is known, where the price is a random variable with a known probability distribution. The firm chooses output to maximize its expected utility. It is well-known that, in the presence of uncertainty, the optimal output of the risk-averse firm is less than that of the risk-neutral firm; moreover, increases in risk aversion, in the sense of Arrow and John Pratt, lead to further diminutions in output. On the other hand, for a fixed degree of risk aversion, the change in output induced by a mean-preserving increase in risk depends on the shape of the cost curve as well as the sign of the third derivative of the utility function u and the sign of the second derivative of $qu'(q)$. Next, competitive industry behavior under uncertainty is analyzed. In order to isolate the effect of uncertainty, firms are assumed to be risk neutral. We show that both the optimal number of firms in the industry and excess capacity increase as industry output becomes riskier. Results like this are important for probabilistic economics, for it would be unfortunate if the vitality of the stochastic theory of the firm relied solely on the controversial assumption of risk aversion.[3]

I. Competitive Production for Risk-Averse Firms

We begin by determining the optimal output for a competitive firm with risk aversion. The price $P \geqslant 0$ is a nondegenerate random variable with known distribution function F and mean μ, and the firm has no control whatsoever over it. Accordingly, as this is a competitive environment, it sells all of its output q at the going price p. The value p of P is made known after the firm has decided upon its production quantity q. Ours is a one-period model, so no storage is permitted from one selling period to the next.

The relationship between the firm's profit π and its output q is given by

$$\pi(q) = Pq - C(q) \tag{1}$$

*University of California-Los Angeles. This research was partially supported by the National Science Foundation through grant SOC-7808985.

[1] For an elaboration of this observation, see Kenneth Arrow and our paper.

[2] This theory has been developed by David Baron (1970, 1971, 1973); Ira Horowitz; Hayne Leland; McCall; Edwin Mills (1959, 1962), Rudolph Penner; Agnar Sandmo; Eytan Sheshinski and Jacques Drèze. The multiperiod case has been investigated by Edward Żabel (1967, 1971).

[3] Our paper with Wayne Winston indicates that the empirical resolution of this controversy will be especially difficult.

where $C(q)$, the total cost of producing q units of the product, consists of a fixed cost B and a variable cost $c(q)$. Naturally C is an increasing function. Because we assume that the marginal cost of production is nondecreasing (so C' is positive and nondecreasing), π is a concave function of q. The firm has a strictly concave utility function u and it seeks q to maximize the expected utility U of profits. Thus, the firm seeks the optimal production level q^* where q^* satisfies[4]

$$(2) \qquad U(q^*)=\max_{q>0} U(q)$$

$$(3) \qquad U(q)=Eu(\pi(q))$$

From the first-order condition[5]

$$(4) \quad E\{u'(\pi(q))P\}=C'(q)Eu'(\pi(q))$$

it is not difficult to demonstrate (see our earlier paper) that

$$(5) \qquad C'(q^*)<\mu$$

In view of (5), C' nondecreasing, and the fact that the optimal output either for a risk-neutral firm or for deterministic demand has marginal cost equal to expected price, it follows that *uncertainty combined with risk aversion leads to decreased output.*[6] (If the firm were risk preferent, i.e., u is convex, then the same argument reveals that output increases.)

The issue we address is how mean-preserving increases in the riskiness of P affect output. We utilize both r_u, the Arrow-Pratt measure of absolute risk aversion, as well as their measure R_u of relative risk aversion, where they are given by

$$(6) \qquad r_u(t)=-\frac{u''(t)}{u'(t)}$$

$$(7) \qquad R_u(t)=-\frac{tu''(t)}{u'(t)}$$

We also make use of the following result.[7]

LEMMA 1: *Let X and Z be nonnegative random variables with cumulative distribution functions F and G, respectively. Suppose that Z is riskier than X in the sense of second-order stochastic dominance; i.e.,*

$$\int_0^x F(t)\,dt \leqslant \int_0^x G(t)\,dt \qquad all\ x \geqslant 0$$

Then $Eu(X)\geqslant Eu(Z)$, for all concave functions u provided $E(Z)=E(X)$.

Because a mean-preserving increase in the riskiness of P leaves μ unchanged, it comes as no surprise that such a change has an impact upon q^*. While an increase in risk can lead as anticipated to a decrease in production, increases in output are also possible. As is true in the study of optimal consumption strategies,[8] the function f defined by $f(t)=tu'(t), t\geqslant 0$, as well as the sign of u''' play a role; in addition, the presence of a fixed cost enters.

THEOREM 1: *Let q_i denote the optimal output when the price P has the same distribution as the random variable P_i, $i=1, 2$, and suppose that P_1 is strictly riskier than P_2 with $E(P_1)=E(P_2)$. Output decreases (i.e., $q_1<q_2$) if f is concave and either u' is convex and marginal cost exceeds average cost or u' is concave and average cost exceeds marginal cost. Output increases (i.e., $q_1>q_2$) if f is convex and either u' is convex and average cost exceeds marginal cost or u' is concave and marginal cost exceeds average cost.*[9]

PROOF:

Let U_i play the role of U in (3) when $P=P_i$ and write E_i to indicate that the expectation is with respect to P_i, $i=1,2$. By

[4] If $E\pi(q)<-B$ all $q>0$, then by Jensen's inequality $U(q)=Eu(\pi(q))<u(E\pi(q))\leqslant u(-B)=U(0)$ so that $q^*=0$. To avoid trivialities, assume $E\pi(q)>-B$ for some q. Because C' is nondecreasing, this means that $EP>C'(0)$.

[5] The concavity of u and π guarantees that of U.

[6] Sandmo was the first to obtain this result.

[7] Lemma 1 is well known and dates back to the 1920's (see David Schmeidler). In particular, the provision $E(Z)=E(X)$ enables us to remove the restriction that u be increasing.

[8] See Leonard Mirman and our paper.

[9] In order to obtain strict inequality between q_1 and q_2, the concavity or convexity of f must be strict or it must be strict for u' and the difference between marginal and average cost must be nonzero.

hypothesis $EP_1 = EP_2$ and f is concave, so applying Lemma 1 to f yields

$$\begin{aligned} U_1'(q) - U_2'(q) &= E_1\{Pu'(\pi)\} \\ - E_2\{Pu'(\pi)\} &- C'(q)\left[E_1u'(\pi) - E_2u'(\pi)\right] \\ &= \frac{1}{q}\left[E_1f(\pi) - E_2f(\pi)\right] \\ + \left[\frac{C(q)}{q} - C'(q)\right] &\left[E_1u'(\pi) - E_2u'(\pi)\right] \\ \leqslant \left[\frac{C(q)}{q} - C'(q)\right] &\left[E_1u'(\pi) - E_2u'(\pi)\right] \end{aligned}$$

Because marginal cost exceeds average cost, $C(q)/q < C'(q)$. Applying Lemma 1 to the convex function u', we obtain $U_1'(q) < U_2'(q)$, whence $q_1 < q_2$. The other results are obtained in the same manner.

The most familiar utility functions satisfying f concave and u' convex are the iso elastic utility functions (i.e., the ones exhibiting constant relative risk aversion) with parameter $\gamma > 0$, i.e., $u(x) = x^{\gamma}/\gamma$, $0 < \gamma < 1$; f and u' are convex if $\gamma < 0$; the case $u(x) = lnx$—the equivalent of $\gamma = 0$—has u' strictly convex and $f(x) \equiv 1$. In addition, the utility functions $u(x) = -e^{-\lambda x}$, $\lambda > 0$, with constant absolute risk aversion, have strictly convex derivatives and f is convex if $R_u(x) = x \geqslant 2$ and concave if $R_u(x) \leqslant 2$.

In the short run, the assumption that marginal cost exceeds average cost means that there are no fixed costs associated with the fixed factors; on the other hand, if average cost exceeds marginal cost, then the fixed factors do impose costs, but they do not cause average variable costs to rise in the region of optimal short-run output. Of course, a short-run U-shaped average cost curve is most realistic.

As presented, Theorem 1 does not cover the case of a U-shaped average cost curve. To do so, let $\bar{q}$ denote the point at which average cost is minimized,[10] $0 < \bar{q} < \infty$. The direction of change is as per Theorem 1 if we replace the phrase "marginal cost exceeds average cost" by $q_2 > \bar{q}$ and "average cost exceeds marginal cost" by $q_2 < \bar{q}$.[11] While it is easy to provide reasonable examples (of C and u) wherein $q_2 < \bar{q}$, the assumption $q_2 > \bar{q}$ is preferable in that the expected profit $E\pi(q)$ increases with output on the interval $[0, q_\mu]$ whenever $\mu > C(\bar{q})/\bar{q}$, where $C'(q_\mu) = \mu$. Furthermore, u' is convex for all parameter values in both examples above. Hence, we might view the case f concave, u' convex, and $q_2 > \bar{q}$ as "typical." In this sense, intuition is validated as output does indeed decrease with a mean-preserving increase in risk.

In studying the long run, suppose there are two factors, y_1 and y_2, and the production function h is homogeneous of degree s, i.e., $q = h(ly_1, ly_2) = l^s h(y_1, y_2)$. Suppose further that (y_1^*, y_2^*) is the optimal combination of y_1 and y_2 to produce one unit of q when the factor prices are w_1 and w_2, respectively. Then, $q = h(ly_1^*, ly_2^*) = l^s$, and total cost is given by $C(q) = (w_1y_1^* + w_2y_2^*)l \equiv \alpha l$. Thus, the long-run total cost function is given by $C(q) = \alpha q^{1/s}$.

When $s = 1$ (constant returns to scale), $q_1 < q_2$ provided f is concave, whereas $q_1 > q_2$ when f is convex. When $s < 1$ (decreasing returns to scale), marginal cost exceeds average cost and $q_1 < q_2$, provided f is concave and u' is convex. The inequality is reversed when f is convex and u' is concave. In particular, note that increased risk leads to a diminution in output if 1) entrepreneurs have constant absolute risk aversion and relative risk aversion is less than or equal to

[10] Presumably, $\mu > C(\bar{q})/\bar{q}$.

[11] Yasunori Ishii showed that $q_1 < q_2$ if r_u is nonincreasing and $P_1 = \delta P_2 - (\delta - 1)EP_2$ with $\delta > 1$. Our Theorem 1 is more general in that we permit P_1 to be any mean-preserving spread and not merely one that changes the scale of P_2. In addition, Theorem 1 covers many cases in which r_u is not nonincreasing. The price of Theorem 1's added generality is the restrictions placed upon either $C(q)/q - C'(q)$ or $q_2 - \bar{q}$. For instance, Theorem 1 does not apply to the case where $u(t) = ln(a+t)$—so u' is strictly convex and f is strictly concave—and $C(q) = B + cq$ with $a > 0$ and $c > 0$ unless $B = 0$. (If $a = 0$—in which case f is also concave—and $B > 0$, it *appears* that Theorem 1 and Ishii's result are in conflict; but in fact $\max_{q>0}|E_i ln(\pi(q))| = \infty$, resolving the apparent conflict.)

two, and 2) returns to scale are nonincreasing.

When $s>1$ (increasing returns to scale), marginal cost is less than average cost and $q_1<q_2$, provided f is concave and u' is concave. The inequality is reversed when f is convex and u' is convex. (In this case, π is no longer concave; consequently we must assume there is a unique solution to (4).)

If, as is likely for a competitive industry, the long-run average cost curve is U-shaped, then Theorem 1 applies just as it did for the short-run analysis.

II. The Competitive Industry under Uncertainty

In their recent paper, Sheshinski and Drèze studied competitive industry behavior under uncertainty. The industry is composed of s identical firms producing a single product. Industry demand is a nondegenerate random variable $Q\geqslant 0$ with $EQ=\mu>0$; thus, the price elasticity is zero. After Q has been observed, the demand is divided equally[12] amongst the s firms. As there is no storage, each firm produces Q/s. The total cost, average cost, and marginal cost associated with a firm's producing q units is denoted by $T(q)$, $A(q)$, and $M(q)$, respectively. They assume (a) there are no barriers to entry, (b) the average cost curve is U-shaped and continuously differentiable, and (c) the marginal cost is increasing and convex. As the industry is competitive, the price p will be equal to the firm's marginal cost.

Because A is U-shaped, there is a number $\bar{q}$ such that

$$(8)\qquad A'(q)\begin{cases} <0, & \text{for } q<\bar{q} \\ >0, & \text{for } q>\bar{q} \end{cases}$$

and $$M(q)\begin{cases} <A(q), & \text{for } q<\bar{q} \\ >A(q), & \text{for } q>\bar{q} \end{cases}$$

for

$$(9)\qquad M(q)=A(q)+qA'(q)$$

Coupling (9) and the fact that price equals marginal cost reveals that $\pi(q)$, the profit due to the firm's producing q units, satisfies

$$(10)\qquad \pi(q)=qM(q)-T(q)$$
$$=q[M(q)-A(q)]=q^2A'(q)$$

Moreover, (9), (10), and (c) imply that π is increasing and strictly convex.

The first problem is to find the industry size s^* that minimizes the expected cost C for the industry. Thus, the optimal industry size s^* satisfies[13]

$$(11)\qquad C(s^*)=\min_{s>0} C(s)$$

$$(12)\qquad C(s)=E\{sT(Q/s)\}$$

and the first-order condition is

$$(13)\qquad 0=C'(s)=-E\pi(Q/s)$$

The mean output per firm in an industry of optimal size is, of course, $q^*=\mu/s^*$. They demonstrate that uncertainty causes the output q^* for an industry with risk-neutral firms to be smaller than the output $\bar{q}$ that minimizes average cost, whence $A(q^*)>M(q^*)$.

From (13) and the fact that there is free entry, it is clear that the number of firms for which expected profit equals zero is precisely s^*. Thus, the competitive equilibrium is efficient in that total expected cost is at its minimum; however, it is characterized by excess capacity in that $q^*<\bar{q}$.

Sheshinski and Drèze demonstrated that in equilibrium the number of firms increases with an increase in either the mean or the variance of Q. We consider the impact of a mean-preserving increase in the riskiness of Q.

THEOREM 2: *In equilibrium the number s^* of firms increases with mean-preserving increases in the riskiness of Q.*[14]

[12] This is efficient because marginal costs are increasing.

[13] For convenience we do not restrict s to be an integer. Accordingly, we need not have $\mu>\bar{q}$.

[14] In their study of a growing oligopoly, A. Michael Spence and Michael Porter obtain a result analogous to Theorem 2. They conjecture that uncertainty reduces the concentration of a mature industry.

PROOF:

Let Q_2 be riskier than Q_1 in the sense of second-order stochastic dominance with $EQ_2 = EQ_1$ and $Q_1 \neq Q_2$. Denote by s_i the industry size that minimizes the expected cost $C_i(\cdot)$ for the industry when the total demand is Q_i, $i=1,2$. Because π is strictly convex it follows from Lemma 1 that $E\pi(Q_2/s) > E\pi(Q_1/s)$. Thus

$$C_2'(s) < C_1'(s) \quad s \geqslant 0$$

and, in particular,

$$C_2'(s_1) < C_1'(s_1) = 0 \tag{14}$$

Joining (14) and the fact that C_2' is an increasing function (it is also concave) establishes $s_2 > s_1$.

An immediate corollary of Theorem 2 is that q^*, the mean output per firm in an industry of optimal size, decreases with mean-preserving increases in risk as $q^* = \mu/s^*$. Thus, the excess capacity $\bar{q} - q^*$ increases as risk increases.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1971.

D. P. Baron, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.*, Oct. 1970, *11*, 463–80.

______, "Demand Uncertainty in Imperfect Competition," *Int. Econ. Rev.*, June 1971, *12*, 196–208.

______, "Point Estimation and Risk Preferences," *J. Amer. Statist. Assn.*, Dec. 1973, *68*, 944–50.

Ira Horowitz, *Decision Making and the Theory of the Firm*, New York 1970.

Y. Ishii, "On the Theory of the Competitive Firm Under Price Uncertainty: Note," *Amer. Econ. Rev.*, Sept. 1977, *67*, 768–69.

H. Leland, "Theory of the Firm Facing Random Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

S. A. Lippman and J. J. McCall, "The Economics of Uncertainty: Selected Topics and Probabilistic Methods," in Kenneth J. Arrow and Michael Intriligator, eds., *Handbook of Mathematical Economics*, Amsterdam 1981.

______, ______ **and W. Winston,** "Risk Aversion, Bankruptcy, and Wealth Dependent Decisions," *J. Bus., Univ. Chicago*, July 1980, *58*, 285–96.

J. J. McCall, "Competitive Production for Constant Risk Utility Functions," *Rev. Econ. Studies*, Oct. 1967, *34*, 417–420.

Edwin S. Mills, "Uncertainty and Price Theory," *Quart. J. Econ.*, Feb. 1959, *73*, 116–30.

______, *Price, Output, and Inventory Policy*, New York 1962.

L. Mirman, "Uncertainty and Optimal Consumption Decisions," *Econometrica*, Jan. 1971, *39*, 179–85.

R. G. Penner, "Uncertainty and the Short-Run Shifting of the Corporation Tax," *Oxford Econ. Papers*, Mar. 1967, *19*, 99–110.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan. 1964, *32*, 122–36.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

D. Schmeidler, "A Biographical Note on a Theorem of Hardy, Littlewood, and Polya," *J. Econ. Theory*, Feb. 1979, *20*, 125–28.

E. Sheshinski and J. H. Drèze, "Demand Fluctuations, Capacity Utilization, and Costs," *Amer. Econ. Rev.*, Dec 1976, *66*, 243–47.

A. M. Spence and M. E. Porter, "The Capacity Expansion Process in a Growing Oligopoly: The Case of Corn Wet Milling," in John J. McCall ed., *The Economics of Information and Uncertainty*, Chicago 1981.

E. Zabel, "A Dynamic Model of the Competitive Firm," *Int. Econ. Rev.*, June 1967, *8*, 194–208.

______, "Risk and the Competitive Firm," *J. Econ. Theory*, June 1971, *3*, 109–33.

Price Regulation, Product Quality, and Asymmetric Information

By DAVID P. BARON*

Direct economic regulation of prices at the firm or industry level involves rules that specify prices based on the costs or profits of those firms or on exogenous factors affecting their performance. These rules may be explicit functions, such as fuel adjustment clauses that automatically adjust electricity rates in response to changes in fuel costs,[1] or they may be implicit rules such as those that yield prices based on estimates of "test year" costs and "revenue requirements."[2] A common characteristic of these price-setting procedures is that prices are to some extent based on the costs actually incurred by the firms being regulated, and when those costs depend on the decisions of the firms, an incentive problem arises. The usual response to such incentive problems is to create mechanisms that eliminate or at least lessen their consequences.[3] This paper instead indicates that the incentive problem created by a cost-based pricing rule can be used constructively to achieve welfare gains when output has a quality dimension and the regulator lacks the statutory authority and the information to implement a first best policy.

The model to be considered pertains to an individual firm or an industry that is subject to price regulation but is able to choose the quality of its output. Given a regulated price, firms will undersupply product quality relative to the socially optimal level. While this undersupply could be overcome through the direct regulation of quality, a regulatory commission may not have the statutory authority to do so. Even if the regulator had the requisite authority, it may not have the cost information that would be required for optimal regulation, since firms are likely to be better informed about costs than is a regulator. A regulatory policy formulated to deal with the undersupply of quality therefore must be based on limit and asymmetric information. The approach to regulation taken here is to delegate the choice of quality to the firms while providing incentives to increase its supply through the relationship set between price and cost. For example, by setting price as an increasing function of cost, the supply of product quality can be stimulated at the expense of giving firms some degree of control over price.[4]

*J. K. Kellogg Graduate School of Management, Northwestern University. I would like to thank Bengt Holmstrom and Raymond De Bondt for their comments. The research has been supported by National Science Foundation Grant SOC 77-07251.

[1] Raymond De Bondt and I (1980a, b) have analyzed the incentive problems associated with average cost pass throughs and have characterized the optimal pass-through function when a regulator has a limited ability to monitor the performance of a firm.

[2] Robert Taggart and I considered the incentive problems inherent in such procedures.

[3] See my paper with Roger Myerson, for example.

[4] An example of the type of situation to which the model considered here is relevant is the system of price regulation practiced in Belgium, which since 1950 has had a form of price regulation for virtually its entire economy. One instrument used in this regulation is a "price calculation contract" under which "output prices are determined by a formula based on the cost of production, and price changes occur automatically in response to changes in cost..." (Baron and De Bondt, 1980a, p. 72). A number of industries supplying such products as petroleum, bread, chocolate, electric home appliances, imported wood, composite animal feed, margarine, and nonferrous metals have operated under such a contract. Firms in these industries have varying degrees of control over the quality of the products they provide and thus have some opportunity to influence price. For example, the contract for the cattle feed industry constrains the mix of the feeds in a blend in order to limit the ability of firms to manipulate the quality and hence cost and price. In the United States, prices apparently have been set as a function of costs through the revenue requirements approach for public utility pricing. For example, a telephone company can determine the reliability and durability of the terminal equipment it supplies and can choose the probability of obtaining a dial tone within a particular time interval. An electric utility can determine its capacity and hence its ability to meet peak demands without voltage reduc-

For the example to be presented in Section III, a regulatory policy that sets price as a markup above marginal costs yields the socially optimal price and quality that the regulator would implement if it had the authority and the information needed to regulate both price and quality. In general, however, the socially optimal outcome cannot be attained through cost-based price regulation when there is an informational asymmetry. If an additional regulatory instrument analogous to Martin Weitzman's quantities or targets can be utilized, however, the socially optimal solution can be achieved. Furthermore, the optimal cost target has the property that the expected target payment is zero and hence no *ex ante* transfer between consumers and firms is made.

I. The Model and the Full-Information Solution

When price is regulated, firms may be able to compete through design and performance features, durability, convenience of use, reliability, etc., which collectively will be referred to as "quality." In order to simplify the analysis, each firm will be assumed to produce only one quality level although a more general formulation would allow firms to produce a number of models with different quality levels. One explanation of the supply of a single model by each firm is that there are fixed costs associated with the number of models produced, and hence a firm finds it optimal to produce only one.[5] Similarly, the regulatory authority will be assumed to set a single price for the product in question, and given that price, firms choose the level of quality they wish to supply. To facilitate the analysis and to obtain tractable results, however, the demand functions of individual firms will be assumed to be symmetric in the quality supplied by the firms in the industry, and hence, there will be an equilibrium in which every firm finds it optimal to supply the same level of quality. This corresponds to an industry in which firms have access to the same technology and hence can replicate the quality of the product supplied by any other firm.

The demand q_i for the output of firm i is given by the function

$$q_i = q_i(p, r_1, \ldots, r_n)$$

which is assumed to be decreasing in the regulated price p and increasing, concave, and symmetric in the product quality r_i, $i=1,\ldots,n$. The production technology of the firm will be assumed to be characterized by constant returns to scale with an exogenous marginal cost c for a "basic" model ($r_i = 0$). The additional marginal cost of producing a product of quality r_i will be denoted by $g(r_i)$, which is assumed to be strictly increasing and convex with $g(0)=0$. The marginal cost y_i of supplying a unit with quality level r_i thus is $y_i = c + g(r_i)$.

The firm is also assumed to incur a fixed cost K associated with the supply of a model, so profit π^i is

$$\pi^i = (p - y_i)q_i - K = (p - c - g(r_i))q_i - K$$

The profit function π^i will be assumed to be concave in (p, r_i) for given $(r_1,\ldots,r_{i-1}, r_{i+1},\ldots,r_n)$ and the required differentiability will be assumed. Each firm is assumed to behave in a Nash manner and thus maximizes its profit given the quality levels of the other $n-1$ firms. Since the firms in the industry are identical, attention will be restricted to a symmetric industry equilibrium.

In practice, a regulatory authority does not have the same information as does the firm, and for the purpose of the analysis here the unit cost c will be considered uncertain and to represent characteristics of production or factor prices that firms will be able to take into account when making their quality decisions but which the regulator will not be able to observe. That is, at the beginning of the period the regulator specifies a price function and the firms decide to incur the fixed cost K based on their common prior information about c, which will

tions. These quality choices affect cost and hence affect the regulated price.

[5] A. Michael Spence (1976) has shown that the presence of fixed costs limits product variety.

be represented by a density function $f(c)$ defined on $c \in [0, \infty)$. The regulator and the firms are assumed to have shared this information and hence to hold the same probability distribution on c. Prior to choosing its product quality, the firm is assumed to learn the cost c and hence to choose quality conditionally on c. A firm thus will agree to supply at the regulated price if its expected profit $\int \pi^i f(c)\,dc$ is at least as great as its reservation level $\pi^*(\pi^* \geqslant 0)$. The regulator does not observe c and thus has less information regarding costs than do firms.[6] The task of the regulator is to choose a regulatory policy that achieves its objectives given this informational asymmetry.

The regulator is assumed to have both consumer's surplus and producer's surplus objectives, which will be represented by the maximization of consumer's surplus subject to a constrained level of producer's surplus. Since the cost c is uncertain at the time the regulatory policy is formulated, expected consumer's surplus will be taken as the measure of consumer's welfare.[7] The results obtained with this formulation can be easily modified to yield the solution that maximizes total surplus.

The full-information optimal regulatory policy is that which the regulatory authority would use if it knew the cost c, and hence that policy consists of a price function $p(c)$ and product quality functions $r_i(c)$, $i = 1, \dots, n$. The regulator's program given full information is thus

$$(1) \qquad \max_{\substack{p(c), r_i(c) \\ i=1,\dots,n}} = \sum_{i=1}^{n} \int S^i(c) f(c)\,dc$$

such that

$$\int \pi^i f(c)\,dc \geqslant \pi^*, \quad i = 1, \dots, n$$

where the conditional consumer's surplus $S^i(c)$ for the output of firm i is given by $S^i(c) = \int_{p(c)}^{\infty} q_i(p^+, r_1(c), \dots, r_n(c))\,dp^+$. Since the full-information policy is conditional on c, the optimal policy is the same as for a deterministic model in which c is known and satisfies the necessary conditions

$$(2) \quad \sum_{i=1}^{n} S_p^i + \sum_{i=1}^{n} \hat{\lambda}^i \pi_p^i = 0$$

$$(3) \qquad S_{r_i}^i + \hat{\lambda}^i \pi_{r_i}^i = 0, \qquad i = 1, \dots, n$$

where $\hat{\lambda}^i$ is the multiplier on the ith constraint in (1). A symmetric policy, $\hat{p}(c)$, $\hat{r}_i(c) = \hat{r}(c)$, $i = 1, \dots, n$, and $\hat{\lambda}^i = \hat{\lambda}$, $i = 1, \dots, n$, will be optimal given the assumptions on demand and costs.

To characterize the full-information policy, (2) and (3) may be written as

$$(2') \quad -q(1-\hat{\lambda}) + \hat{\lambda}(\hat{p}(c) - c - g(\hat{r}(c)))q_p = 0$$

$$(3') \quad \int_{\hat{p}(c)}^{\infty} q_r\,dp^+ + \hat{\lambda}((\hat{p}(c) - c - g(\hat{r}(c)))q_r - g'(\hat{r}(c))q) = 0$$

From (2′) if $\hat{\lambda} < 1$, $\hat{p} - c - g(\hat{r}(c)) < 0$ for all c, and hence expected profit is negative which violates the constraint in (1). Consequently, $\hat{\lambda} \geqslant 1$ and the price $\hat{p}(c)$ is at least as great as marginal cost $y = c + g(\hat{r}(c))$ for all c. If $\pi^* = K = 0$, $\hat{\lambda} = 1$ and price is equal to marginal cost. From (2) the marginal profit π_p^i is positive indicating that regulation is effective in maintaining price below

[6] In an analogous situation, Weitzman states:

> An essential feature of the regulatory environment I am trying to describe is uncertainty about the exact specification of each firm's cost function. In most cases even the managers and engineers most closely associated with production would be unable to precisely specify beforehand the cheapest way of generating various hypothetical output levels. Because they are yet further removed from the production process, the regulators are likely to be vaguer still about a firm's cost function. [p. 684]

[7] The income elasticity is assumed to be zero, and because of the uncertainty about c, expected consumer's surplus will be used under the assumption that consumers are risk neutral. Without the risk-neutrality assumption, expected consumer's surplus is only an approximation of consumer welfare as analyzed by Richard Schmalensee (1972).

that which the firm would prefer given $\hat{r}(c)$. Without further specification of the demand function the properties of $\hat{p}(c)$, $\hat{r}(c)$, and $\hat{y}(c)$ cannot be established unambiguously. It will be assumed that marginal cost $\hat{y}(c)$ is increasing in c, which seems reasonable and is the case for the example of Section III.

In the regulatory framework to be considered here, the regulator does not have the statutory authority to directly regulate quality or to indirectly regulate quality through the imposition of penalties on the firm if it does not supply the desired level of quality.[8] As the following proposition indicates, even if the regulator knew c, the industry would undersupply product quality given a regulated price.

PROPOSITION 1: *Given a regulated price p, the industry undersupplies quality relative to the socially preferred level.*[9]

PROOF:

For a given c the constrained variation $(\partial L/\partial r_i)$ in welfare is from (3)

$$\frac{\partial L}{\partial r_i}=S^i_{r_i}+\lambda_i\pi^i_{r_i}=\int_p^{\infty}q^i_{r_i}dp^+ +\lambda_i\pi^i_{r_i}$$

Given p, the response function $r_i^*(c)$ chosen by the firm satisfies $\pi^i_{r_i}=0$, so evaluating $\partial L/\partial r_i$ at $r_i^*(c)$ yields

$$\left.\frac{\partial L}{\partial r_i}\right|_{r_i=r_i^*(c)}=\int_p^{\infty}q^i_{r_i}dp^+>0 \qquad \text{for all } c$$

Consequently, a greater supply of quality is socially preferred.

The undersupply of quality results because there is no market price that measures the marginal consumer's surplus with respect to quality.

II. Second Best Price Regulation

When the regulator does not know the exogenous component c of cost and cannot regulate quality, the regulator can attempt to alleviate the undersupply of quality through the relationship it sets between price and cost. The approach taken here is to delegate the choice of the regulated price to the firm by choosing a price schedule $p_0(y)$ and allowing the firm to choose its marginal cost and hence the price. Given the price schedule, the firm will choose a quality response function $r(c)$ or equivalently a marginal cost response function $y(c)=c+g(r(c))$. The price p at which the firm's output is sold is then given by $p=p_0(y(c))$. The regulator is assumed to be able to observe y *ex post* and hence to ensure that the firm implements the price corresponding to its marginal cost according to the delegated price schedule. The class of marginal cost response functions $y(c)=c+g(r(c))$ that the regulator can induce by using a price schedule $p_0(y)$ will first be characterized and then an example will be presented in Section III in which the full-information solution can be attained.

To characterize the class of response functions that can be induced by a price function $p_0(y)$, a firm will be viewed as choosing its marginal cost y_i. Letting $h(y_i-c)\equiv g^{-1}(y_i-c)$, profit is

$$(4)\quad \pi^i(p_0(y_i),y_i,c)$$
$$=(p_0(y_i)-y_i)q_i(p_0(y_i),$$
$$h(y_1-c),\ldots,h(y_n-c))-K$$

[8] If the regulator has the requisite authority to impose unlimited penalties on the firm and is able to observe any two of q, c, y, and r, the full-information solution can be achieved by forcing the firm to adopt $\hat{r}(c)$ when the price $\hat{p}(c)$ is set. For example, if q and y can be observed, r and c can be deduced from the demand function and the relationship $y=c+g(r)$. The regulator may then penalize the firms if the quality it chooses does not equal $\hat{r}(c)$. To achieve the full-information price, the regulator can use the price function $\bar{p}(y)\equiv\hat{p}(\hat{y}^{-1}(y))$, where $\hat{y}^{-1}(y)$ is obtained by inverting the relationship $y=\hat{y}(c)\equiv c+\hat{r}(c)$. The authority to impose such penalties will not be assumed to be granted here.

[9] This result is not in conflict with Spence's (1975) Proposition 1 that a monopolist may either under- or oversupply quality relative to the first best solution, since that proposition pertains to a monopolist that is able to choose his price and quality while the result here provides a comparison for a fixed price. Lawrence White has obtained a similar result in the context of a model representing the supply of transportation.

Maximizing with respect to y_i for $y_i \geqslant c$ yields the firm's optimal response function $\bar{y}(c)$, which satisfies the following first-order condition evaluated at a symmetric equilibrium

$$\pi_p p_0' + \pi_y = 0 \tag{5}$$

where the superscript i has been dropped, since no ambiguity will result. To investigate the second-order condition, differentiate (5) with respect to c to obtain

$$\frac{\partial}{\partial y}(\pi_p p_0' + \pi_y)\bar{y}'(c) + \frac{\partial}{\partial c}(\pi_p p_0' + \pi_y) = 0 \tag{6}$$

The last term may be written as[10]

$$\frac{\partial}{\partial c}(\pi_p p_0' + \pi_y) = \pi_p \frac{\partial}{\partial c}\left(\frac{\pi_y}{\pi_p}\right)$$

If regulation is effective in maintaining the price below that which the firm would prefer given its marginal cost $\bar{y}(c)$, the term π_p is positive. Consequently, if $\partial(\pi_y/\pi_p)/\partial c$ is positive and if $\bar{y}'(c)$ is positive as would be expected, the response function $\bar{y}(c)$ is at least a local optimum for the firm. Given that $\partial(\pi_y/\pi_p)/\partial c > 0$, a theorem due to Bengt Holmstrom may be used to establish the following result, the proof of which parallels that of Proposition 4 in Section IV and hence is omitted here.

PROPOSITION 2: *If* $\partial(\pi_y/\pi_p)/\partial c > 0$, *any nondecreasing response function* $\bar{y}(c)$ *satisfying* (5) *is attainable using a price function* $p_0(y)$ *and is a global solution to the firm's problem. Conversely, any response function satisfying* (5) *is nondecreasing.*

The condition $\partial(\pi_y/\pi_p)/\partial c > 0$ states that the ratio of the marginal profit from an increase in marginal cost to the marginal profit from an increase in price is increasing in the exogeneous cost c.

The significance of Proposition 2 is that, by using cost-based price regulation, the regulator can induce the firm to choose any desired quality response function $r(c)$ such that marginal cost $y(c)$ is increasing in c. In particular, if the full-information optimal marginal cost $\hat{y}(c) = c + g(\hat{r}(c))$ is an increasing function, cost-based price regulation can achieve the full-information optimal quality. The resulting price $p_0(\hat{y}(c))$ will not in general equal the full-information optimal price, however. To investigate the condition of Proposition 2 and the properties of the full-information solution, the demand function and the quality cost function will be further specified.

III. An Example

As an example, let the demand function q_i be specified as[11]

$$q_i = \frac{r_i^\eta/n}{\left(\sum_j r_j^\eta/n\right)}\left(\sum_j r_j/n\right)^\alpha G(p)$$

$$0 < \alpha \leqslant 1, \quad 0 < \eta \leqslant 1, \quad G'(p) < 0$$

so that industry demand is an increasing function of the average product quality $\Sigma r_j/n$, and the demand for firm i's product is an increasing function of its relative quality as measured by $r_i^\eta/(\Sigma r_j^\eta/n)$. To facilitate the analysis, the quality cost function will be specified as $g(r_i) = r_i^\beta$, $\beta \geqslant 1$. The full-information solution for this specification satisfies the necessary conditions

$$-(1-\hat{\lambda})G(\hat{p}) + \hat{\lambda}(\hat{p} - c - \hat{r}^\beta)G'(\hat{p}) = 0 \tag{7}$$

$$\gamma \int_{\hat{p}}^{\infty} G(p^+)\,dp^+ + \hat{\lambda}(-\beta\hat{r}^\beta + (\hat{p} - c - \hat{r}^\beta)\gamma)G(\hat{p}) = 0 \tag{8}$$

where $\gamma = \eta(n-1)/n + \alpha/n$ is the quality

[10] John Riley has presented a similar analysis.

[11] This formulation is similar to that used by Schmalensee (1977). The demand function may be interpreted as an industry demand

$$Q(p, r_1, \ldots, r_n) = \left(\Sigma r_j/n\right)^\alpha G(p)$$

with the market share s_i obtained by firm i given by

$$s_i = \frac{1}{n}\frac{r_i^\eta}{\left(\Sigma r_j^\eta/n\right)}$$

elasticity of a firm's demand q_i. The full-information price function $\hat{p}(c)$ and the marginal cost $\hat{y}(c)=c+\hat{r}(c)^{\beta}$ are increasing functions of c, but the product quality $\hat{r}(c)$ may increase or decrease with c.[12] An increase in the cost of a basic model thus may cause firms to increase or decrease their product quality, but in either case the result is an increase in the marginal cost and the price of the model supplied.[13]

To stimulate the supply of product quality, the regulator can base price $p_0(y)$ on marginal cost. The condition in (5) is

$$\begin{aligned}&(G(p_0)\\&\quad+(p_0-\bar{y})G'(p_0))(\bar{y}-c)^{a/\beta}n^{-1}p_0'(\bar{y})\\&\quad+\left[-\beta(\bar{y}-c)+(p_0-\bar{y})\gamma\right]\\&\qquad\qquad(\bar{y}-c)^{\alpha/\beta-1}G(p_0)n^{-1}\beta^{-1}=0\end{aligned}$$

Evaluating the condition in Proposition 2 yields

$$\begin{aligned}&\pi_p\frac{\partial}{\partial c}\left(\frac{\pi_y}{\pi_p}\right)\\&=(p_0-\bar{y})\gamma(\bar{y}-c)^{\alpha/\beta-2}G(p_0)n^{-1}\beta^{-1}>0\end{aligned}$$

so any increasing response function can be attained through price regulation. Since the full-information solution $\hat{y}(c)$ is an increasing function, it can be attained, but in general the resulting price $p_0(\hat{y}(c))$ will not be full-information optimal. A further specification of the demand function, however, yields a case in which the full-information solution is attained.

If the function $G(p)$ is specified as $G(p)=p^{\varepsilon}$, $\varepsilon<-1-\gamma/\beta$, the full-information solution is

$$(9)\qquad \hat{p}(c)=\frac{\hat{\lambda}\varepsilon\left(c+\hat{r}(c)^{\beta}\right)}{\hat{\lambda}(\varepsilon+1)-1}$$

$$(10)\qquad \hat{r}(c)=\left[\frac{-\gamma c}{\beta(\varepsilon+1)+\gamma}\right]^{1/\beta}$$

The undersupply of product quality can be overcome for this example by using a simple markup pricing rule of the form

$$(11)\quad p_0(y_i)=y_i(1+m)=\left(c+r_i^{\beta}\right)(1+m)$$

where m is the percentage markup above marginal cost. The equilibrium quality $r_m(c)$ given the price function in (11) is

$$(12)\qquad r_m(c)=\left[\frac{-\gamma c}{\beta(\varepsilon+1)+\gamma}\right]^{1/\beta}$$

which is independent of m and equals the full-information quality $\hat{r}(c)$.

The incentive created by markup regulation thus leads the industry in this case to supply the optimal product quality. The resulting price $p_m(c)$ given (11) and (12) satis-

[12] Total differentiation of (7) and (8) yields

$$\hat{p}'(c)=-\beta^2\hat{\lambda}^2G''(\hat{p})\hat{r}^{\beta-1}/D>0$$

$$\begin{aligned}&\hat{y}'(c)=\beta^2\hat{r}^{\beta-1}\hat{\lambda}G(\hat{p})\left[(1-2\hat{\lambda})G'(\hat{p})-\hat{\lambda}G''(\hat{p})\right]/D\\&\qquad>0\end{aligned}$$

where $D>0$ is the determinant of the matrix of second partial derivatives and the term $[(1-2\hat{\lambda})G'(\hat{p})-\hat{\lambda}G''(\hat{p})]$ is positive from the second order conditions.

[13] Given a fixed regulated price p, a firm's optimal quality response function $r^*(c)$ expressed as a function of c is, at a symmetric equilibrium,

$$r^*(c)=\left(\frac{\gamma(p-c)}{\beta+\gamma}\right)^{1/\beta}$$

The full-information optimal response function, given the full-information optimal $\hat{p}(c)$, is from (8)

$$\begin{aligned}\hat{r}(c)=\Bigg[\gamma\Bigg(\hat{p}(c)-c+\left(\int_{\hat{p}(c)}^{\infty}G(p^+)\,dp^+\right)\\/\left(\hat{\lambda}G(\hat{p}(c))\right)\Bigg)^{1/\beta}\Bigg]/(\beta+\gamma)\end{aligned}$$

Comparing the full-information response function $\hat{r}(c)$ and the response function $r^*(c)$ indicates that

$$\hat{r}(c)^{\beta}-r^*(c)^{\beta}=\left(\frac{1}{\beta+\gamma}\right)\frac{\int_{\hat{p}(c)}^{\infty}G(p^+)\,dp^+}{\hat{\lambda}G(\hat{p}(c))}>0$$

Consequently, if a regulator were to implement the full-information optimal price function $\hat{p}(c)$ and did not have the authority to regulate quality, the industry will undersupply product quality as indicated in Proposition 1.

fies

$$(13)\qquad p_m(c) = (1+m)\frac{\beta c(\varepsilon+1)}{\beta(\varepsilon+1)+\gamma}$$

which is equal to the full-information optimal price in (9) if

$$(14)\qquad m = \frac{1-\hat{\lambda}}{\hat{\lambda}(\varepsilon+1)-1}$$

This result is stated as

PROPOSITION 3: *For the example with a constant price elasticity demand function with* $\varepsilon < -1-\gamma/\beta$, *markup regulation of the form in* (11) *with markup given in* (14) *yields the full-information price function* $\hat{p}(c)$ *and quality response function* $\hat{r}(c)$.

This result indicates that the incentive problem created by the dependence of price on a firm's marginal cost can be used to induce the firm to supply more quality than it otherwise would when price is independent of cost. If the markup is chosen correctly, both the full-information price and quality can be attained. In general, however, the full-information solution cannot be attained using only price as a regulatory instrument. In the next section an additional regulatory instrument will be introduced that allows the full-information solution to be attained.

IV. Cost Targets and Welfare Improvements

In the absence of the authority to regulate quality, the regulator may attempt to achieve welfare gains by employing additional regulatory instruments that indirectly affect the supply of quality. Since the regulator is able to observe marginal cost y, the instrument to be considered here is a cost target $T(y)$ that involves a payment from consumers to the firm. This section demonstrates that a cost target $T(y)$ exists that induces the firm to choose the full-information cost response $\hat{y}(c)$ when the regulated price is set according to the full-information price function $\hat{p}(c)$. Letting $c=\hat{y}^{-1}(y)$ be the exogenous cost corresponding to $y=\hat{y}(c)$, the full-information price will be written as a function $\bar{p}(y)$ defined by $\bar{p}(y) \equiv \hat{p}(\hat{y}^{-1}(y))$.

Given $\bar{p}(y)$ and a cost target $T(y)$, a firm's profit is $\bar{\pi}(y,c)+T(y)$, where $\bar{\pi}(y,c) \equiv \pi(\bar{p}(y),y,c)$. The optimal response function $y(c)$ thus satisfies, assuming that $T(y)$ is differentiable,

$$(15)\qquad \bar{\pi}_y + T'(y) = \pi_p \bar{p}'(y) + \pi_y + T'(y) = 0 \qquad \text{for all } c$$

Inspection of (15) indicates that the regulator can choose a cost target $\hat{T}(y)$ such that the full-information marginal cost $\hat{y}(c)$ satisfies the necessary condition for optimality. To show that the regulatory policy $(\bar{p}(y), \hat{T}(y))$ yields the full-information optimum, however, it is necessary to demonstrate that given that policy a firm will find the response function $\hat{y}(c)$ to be a global optimum.

To investigate the optimality of $\hat{y}(c)$, differentiate the first-order condition in (15) with respect to c to obtain

$$(16)\qquad \frac{\partial}{\partial y}(\bar{\pi}_y + \hat{T}'(y))\hat{y}'(c) + \bar{\pi}_{yc} = 0$$

The full-information response function $\hat{y}(c)$ is strictly increasing in c, so if $\bar{\pi}_{yc} \geq 0$ and $\hat{T}(y)$ is chosen such that $\hat{y}(c)$ satisfies (15), $\hat{y}(c)$ will be a local maximum of the firm's problem. From (4) $\bar{\pi}_{yc}$ may be evaluated as

$$(17)\qquad \bar{\pi}_{yc} = \frac{\partial}{\partial c}(\bar{\pi}_y)$$
$$= \frac{\partial}{\partial c}(-q+(\bar{p}-y)q_y) = q_y - (\bar{p}-y)q_{yy} > 0$$

since $q_c = -q_y$, $q_{yc} = -q_{yy}$, $q_y = q_r h' > 0$, and $q_{yy} = q_r h'' + q_{rr}(h')^2 < 0$, because q_i is increasing and concave in r_i, and h is increasing and concave. Consequently, $\hat{y}(c)$ is a local maximum to the firm's problem.

The above analysis demonstrates that a function $\hat{T}(y)$ exists such that $\hat{y}(c)$ satisfies the first-order condition in (15) and yields at least a local maximum of the firm's problem. The following proposition indicates that $\hat{y}(c)$ is a global optimum to the firm's problem. The proof follows that given by Holmstrom.

PROPOSITION 4: *Given the price function $\bar{p}(y)$ and a target function $\hat{T}(y)$ such that the response function $\hat{y}(c)$ satisfies* (15), *$\hat{y}(c)$ is a global optimum to the firm's problem. Thus, the full-information solution is attainable.*

PROOF:

The full-information optimal $\hat{y}(c)$ will be shown to be the global solution to the firm's problem. Suppose that for some $c_1, y_1 = \hat{y}(c_1)$ satisfying (15) is not a global optimum. That is, the slope $-\bar{\pi}_y(y_1, c_1)$ of the indifference curve in the (y, T) plane defined by

$$\bar{\pi}(y, c_1) + T \equiv \bar{\pi}(y_1, c_1) + \hat{T}(y_1)$$

equals the slope $\hat{T}'(y_1)$ at $(y_1, \hat{T}(y_1))$ as indicated by (15). Since y_1 is not a global optimum, the indifference curve must intersect the function $\hat{T}(y)$ at some other point, say y_2.[14] If $y_2 > y_1$, then $\hat{T}'(y_2) > -\bar{\pi}_y(y_2, c_1)$ in order for T and the indifference curve to intersect at $y_2 > y_1$. Corresponding to y_2 is a $c_2 (c_2 > c_1)$ such that $y_2 = \hat{y}(c_2)$ satisfies (15) or

$$\bar{\pi}_y(y_2, c_2) + \hat{T}'(y_2) = 0$$

Consequently, $-\bar{\pi}_y(y_2, c_2) > -\bar{\pi}_y(y_2, c_1)$

This however contradicts the condition in (17) that $\bar{\pi}_{yc} > 0$. An analogous argument holds if $y_2 < y_1$. Consequently, $\hat{y}(c)$ is a global optimum to the firm's problem, and the firms will thus choose $\hat{y}(c)$ when the regulator uses the regulatory policy

$$(\bar{p}(y), \hat{T}(y))$$

Proposition 4 indicates that the regulator can attain the full-information solution by employing the policy $(\bar{p}(y), \hat{T}(y))$ and delegating the choice of product quality, and hence of price, to the firms in the industry.

Given $(\bar{p}(y), \hat{T}(y))$, the firm will choose $\hat{y}(c)$ and the resulting expected profit in (4) will equal π^*, so the cost target satisfies $\int \hat{T}(\hat{y}(c)) f(c)\, dc = 0$. Consequently, the use of a target function involves no *ex ante* (expected) transfer from consumers to firms.[15]

The optimal cost target $\hat{T}(y)$ may be determined from (3′) and (15) and satisfies

$$\hat{T}'(y) = \frac{1}{\hat{\lambda}} \int_{\bar{p}(y)}^{\infty} q_y\, dp^+ - \pi_p \bar{p}'(y)$$

$$= \left(\int_{\bar{p}(y)}^{\infty} q_y\, dp^+ - q\bar{p}'(y) \right) \Big/ \hat{\lambda}$$

(using (2′))

where the right side is evaluated at $c = \hat{y}^{-1}(y)$. The marginal target function is thus proportional to the marginal gain to consumers from a variation in quality. The marginal gain is composed of the marginal consumer's surplus less the additional expenditure $q\bar{p}'(y)$ for the product necessitated by the dependence of price on quality through the marginal cost. Consequently, if the gain is positive (negative) for a given c at $\hat{y}(c)$, the target cost is increasing (decreasing) at $\hat{y}(c)$ in order to stimulate (retard) the supply of quality.[16, 17]

V. Conclusions

For the model considered here, a cost-based price function can be used to alleviate

[14] This argument requires that the indifference curve be convex in the (y, T) plane, which will be the case if $\bar{\pi}(y, c)$ is concave in y. The latter follows from the concavity of π^i in r_i.

[15] An issue that has not been addressed here is the further use of a cost target to achieve welfare gains by enabling marginal cost pricing to be used. If lump sum transfers can be made in the full-information case and the social objective is the maximization of total surplus, then $\hat{p}(c) = \hat{y}(c)$ is the optimal price function, and the lump sum payment is equal to $\pi^* + K$. With this policy the only incentive for the firm is provided by the cost target. From (3′) the cost target, evaluated at $c = \hat{y}^{-1}(y)$, satisfies the following conditions: $T'(y) = \int_y^{\infty} q_r\, dp^+ - g'q$ and $\int T(\hat{y}(c)) f(c)\, dc = K + \pi^*$.

[16] This cost target is similar to the regulatory instrument used by Loeb and Magat, and Myerson and myself, although we considered the case in which the firm knows its cost prior to contracting with the regulator and the product did not have a quality dimension.

[17] It is easily verified that for the constant price elasticity example of Section III, $\hat{T}'(y) = 0$ for all y, so no incentive is needed to attain the full-information solution.

the market failure that results in the undersupply of product quality. While a socially optimal regulatory policy can be implemented when the regulator has the same information as does the firm and has the authority to directly regulate quality, the absence of that authority and asymmetric information about cost can limit the attainment of social welfare goals. For the example of Section III, the full-information solution can be attained by setting price as a fixed markup above marginal cost, and the markup can be chosen so that the resulting price is socially optimal. This result does not obtain in general, so cost-based price regulation yields only a second best solution. If, however, a cost target can be used as a regulatory instrument, the socially optimal price and quality can be achieved by specifying a set of prices $\bar{p}(y)$ and cost targets $\hat{T}(y)$, and delegating the choice of cost y, and hence of price, to the firm. This indirect means of regulating quality overcomes both the absence of the authority to directly regulate quality and the asymmetric information about cost.[18]

[18] This regulatory policy, however, is not as straightforward as suggested by this analysis because the firms in this model have no opportunity to produce inefficiently. A policy that bases price on cost may create an incentive for pure waste, as indicated by De Bondt and myself (1980b), and those losses must be weighed against the gains from enhancing the supply of quality.

REFERENCES

D. P. Baron and R. R. De Bondt, "Fuel Adjustment Mechanisms and Economic Efficiency," *J. Ind. Econ.*, Mar. 1979, *27*, 243–61.

______**and R. R. De Bondt,** (1980a) "Belgian Price Regulation and Non-Price Competition," in B. M. Mitchell and P. R. Kleindorfer, eds. *Regulated Industries and Public Enterprise: European and United State Perspectives*, Lexington, Mass. 1980.

______**and**______ (1980b) "On the Design of Regulatory Price Adjustment Mechanisms," *J. Econ. Theory*, forthcoming.

______**and R. B. Myerson,** "Regulating a Monopolist with Unknown Costs," disc. paper no. 412, Center Math. Stud. Econ. Manage. Sci., Northwestern Univ. 1979.

______**and R. A. Taggart,** "Regulatory Pricing Procedures and Economic Incentives," in M. A. Crew, ed., *Issues in Public Utility Pricing Regulation*, Lexington, Mass. 1980.

B. Holmstrom, "On Decentialization, Incentives, and Control," unpublished doctoral dissertation, Stanford Univ. 1977.

M. Loeb and W. A. Magat, "A Decentralized Method for Utility Regulation," *J. Law and Econ.*, Oct. 1979, *22*, 399–404.

J. G. Riley, "Competitive Signalling," *J. Econ. Theory*, Apr. 1975, *10*, 174–86.

R. Schmalensee, "Option Demand and Consumer's Surplus: Valuing Price Changes Under Uncertainty," *Amer. Econ. Rev.*, Dec. 1972, *62*, 813–24.

______, "Comparative Static Properties of Regulated Airline Oligopolies," *Bell J. Econ.*, Autumn 1977, *8*, 565–76.

A. M. Spence, "Monopoly, Quality, and Regulation," *Bell J. Econ.*, Autumn 1975, *6*, 417–29.

______, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Stud.*, June 1976, *43*, 217–35.

M. L. Weitzman, "Optimal Rewards for Economic Regulation," *Amer. Econ. Rev.*, Sept. 1978, *68*, 683–91.

L. J. White, "Quality Variation When Prices are Regulated," *Bell J. Econ.*, Autumn 1972, *3*, 425–36.

Variable Returns to Scale in Production and Patterns of Specialization

By ARVIND PANAGARIYA*

This paper discusses some trade and welfare implications of the two-sector model based on increasing returns to scale (*IRS*) in one industry and decreasing returns to scale (*DRS*) in the other. It has been shown by Horst Herberg and Murray Kemp that, given homothetic production functions with *IRS* in one industry and *DRS* in the other, the production-possibilities frontier (*PPF*) is strictly concave to the origin near the *IRS* axis and strictly convex to the origin near the *DRS* axis. While this result constitutes an important finding and is contrary to the general impression among economists, even twelve years after the publication of Herberg and Kemp's paper, no attempt has been made to analyze its implications for welfare and patterns of specialization.[1]

In this paper, I consider a simple two-commodity model with *IRS* in one industry and *DRS* in the other, and discuss some interesting implications of variable returns to scale (*VRS*). First, I demonstrate that, in this model, if a small, open economy specializes completely in production, it will do so in the *DRS* commodity and not in the *IRS* commodity as is generally believed. Second, if, at a given price ratio, an internal production equilibrium exists, a welfare-maximizing small, open economy will never specialize completely in production even though the *PPF* exhibits decreasing opportunity costs over a part of its range. Third, given output-generated economies and diseconomies of scale, welfare maximization for a small country requires a permanent tax subsidy scheme encouraging expansion of the *IRS* industry and contraction of the *DRS* industry. Finally, in a two-country model with identical tastes and technology across the countries, free trade equilibrium may result in incomplete specialization by the exporter of the *IRS* commodity and complete specialization by the exporter of the *DRS* commodity.

In the course of the analysis, I argue that Tinbergen's formulation of Frank Graham's case for protecting the *IRS* industry is incorrect. I then present a possible reformulation of it. In Section I, the basic production model is introduced; in Section II, the model is analyzed; and in Section III, it is argued that all the results continue to hold in the multifactor case.

I. The Production Model

Throughout this paper, it is assumed that there is only one factor of production, labor. This restriction is imposed for convenience; none of the results depend on it. In fact, as shown later in Section III, all the results continue to hold in the multifactor case.

Assume that the economy under consideration produces two commodities, 1 and 2. Let commodity 1 be subject to *DRS* and commodity 2 to *IRS*. Suppose that economies and diseconomies are external to the firm but internal to the industry so that they are consistent with perfect competition and output is priced at average cost. Also, let the economies and diseconomies be output generated; this ensures that, under perfect competition, production takes place along the

*Assistant professor, University of Maryland. I gratefully acknowledge helpful comments from T. J. Bertrand, William Branson, Jon Eaton, Peter Kenen, John Piderit, Jan Svejnar, Laura Tyson, and Larry E. Westphal. I also wish to thank George Borts and an anonymous referee for many helpful suggestions. Any remaining errors are mine.

[1] It may be of interest to note that there is a strong tendency in the literature to draw the *PPF* strictly convex to the origin near the *IRS* axis and strictly concave to the origin near the *DRS* axis. For example, see Jan Tinbergen (p. 192, Figure 4); Richard Caves (p. 172, Figure 7); Charles Kindleberger (p. 35, Figure 2.7b); C. E. Staley (p. 94, Figure 9-1); Caves and Ronald Jones (p. 107, Figure 6.2); Ingo Walter (p. 107, Figure 7-5).

PPF. Finally, assume that production functions are homogeneous. Denoting the output of commodity i produced by a typical firm j, by X_{ij}, amount of labor employed by firm j in industry i by L_{ij}, total output of commodity i by X_i and total amount of labor employed in industry i by L_i, we can write the production functions of the firm and industry, respectively, as

$$(1a) \qquad X_{ij} = X_i^{\alpha_i} L_{ij} \qquad i = 1,2$$

$$(1b) \qquad X_i = \Sigma_j X_{ij} = L_i^{1/(1-\alpha_i)} = L_i^{\delta_i}$$

For notational convenience, I have substituted $\delta_i \equiv 1/(1-\alpha_i)$ in deriving the final form of equation (1b) from (1a). Note that δ_i (or α_i) is the returns-to-scale parameter. Industry i is subject to *IRS* or *DRS* as $\delta_i \gtrless 1(\alpha_i \gtrless 0)$. Given our assumption that industry 1 is subject to *DRS* and industry 2 to *IRS*, we have $\delta_1 < 1$ and $\delta_2 > 1$ ($\alpha_1 < 0$ and $\alpha_2 > 0$).

Observe that according to equation (1a) the firm's output depends parameterically on the industry output, implying that economies or diseconomies are output generated. Moreover, given the industry output, the firm faces constant returns to scale in production; this indicates the external nature of economies and diseconomies of scale.

Given a fixed endowment of labor L, we may write, as full-employment condition,

$$(2) \qquad L_1 + L_2 = L$$

We can now consider the shape of the *PPF* and the competitive equilibrium on the production side of the economy.

A. *The Production-Possibilities Frontier*

Equations (1b) and (2) define the economy's *PPF*. By straightforward manipulations, it can be shown that the slope and the curvature of the *PPF*, respectively, are given by

$$(3) \qquad \frac{dX_2}{dX_1} = -\frac{\delta_2}{\delta_1} X_1^{-(\delta_1-1)/\delta_1} X_2^{(\delta_2-1)/\delta_2}$$

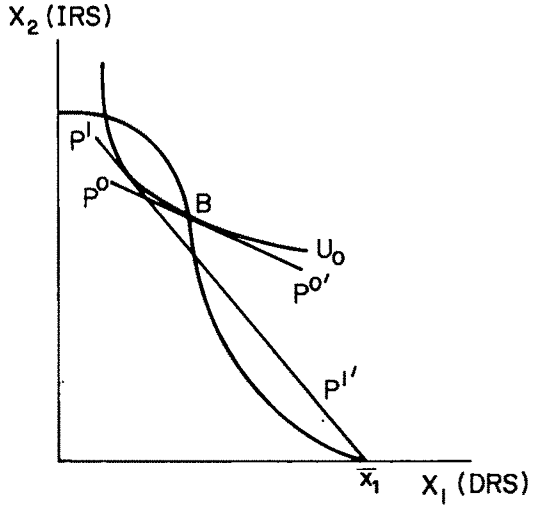


FIGURE 1

$$(4) \qquad \frac{d^2X_2}{dX_1^2} = \frac{1}{\delta_2} X_1^{(1-2\delta_1)/\delta_1} X_2^{(\delta_2-2)/\delta_2} \times \left[(\delta_1-1)X_2^{1/\delta_2} + (\delta_2-1)X_1^{1/\delta_1}\right]$$

By definition, the *PPF* is strictly convex or strictly concave to the origin as $(d^2X_2/dX_1^2) \gtrless 0$. Remembering that $\delta_1 < 1$ and $\delta_2 > 1$, equation (4) implies that $(d^2X_2/dX_1^2) < 0$ in the neighborhood of $X_1 = 0$ and $(d^2X_2/dX_1^2) > 0$ in the neighborhood of $X_2 = 0$. Therefore, the *PPF* is strictly concave to the origin near the $X_2(IRS)$ axis and strictly convex to the origin near the $X_1(DRS)$ axis. This result is the same as that obtained by Herberg and Kemp under the more general assumptions of two factors of production and homothetic production functions. The additional restrictions of homogeneity and one factor of production, imposed in this paper, yield two more properties of the *PPF*. First, setting $(d^2X_2/dX_1^2) = 0$ in equation (4), we obtain a unique point of inflection. Therefore, the *PPF* is shaped like an inverted S as shown in Figure 1. Second, it follows from equation (3) that the slope of the *PPF* approaches zero as output of either commodity approaches zero. This implies that the *PPF* is flat in the neighborhood of zero output of each commodity.

B. *The Competitive Equilibrium*

Under competition, each firm employs labor up to the point where its private value of marginal product equals the wage rate. Therefore, denoting the price of commodity i by p_i and wage rate by w and remembering that the firm's production function is given by equation (1a), we have, as factor market-clearing condition,

$$w = p_1 X_1^{(\delta_1 - 1)/\delta_1} = p_2 X_2^{(\delta_2 - 1)/\delta_2} \tag{5}$$

It immediately follows from equations (3) and (5) that the production equilibrium satisfies the condition[2]

$$\frac{p_1}{p_2} = -\frac{\delta_1}{\delta_2}\frac{dX_2}{dX_1} \tag{6}$$

It is evident from equation (6) that given $\delta_1 < 1$ and $\delta_2 > 1$ the price line is flatter than the *PPF*. That is, as in Figure 1, the price line cuts the *PPF* from below. The intuitive reason for this relationship lies in the nature of returns to scale in the two industries. Given $\delta_1 < 1$, the expansion of a firm in industry 1 generates a negative externality for other firms in that industry. Given $\delta_2 > 1$, a similar change in industry 2 generates a positive externality. As a result, in industry 1, the private marginal product of labor for a firm exceeds the social marginal product, while in industry 2, exactly the opposite holds true. From the point of view of socially optimal resource allocation, this leads to overemployment of labor in industry 1 and underemployment of labor in industry 2. In terms of Figure 1, this is equivalent to the price line being flatter than the *PPF*.

It may be noted that for a given price ratio there exist either two or no internal production equilibria. Beginning from $X_1 = 0$, as we move along the *PPF* towards $X_2 = 0$, the absolute value of the slope of the *PPF* first increases, reaches a maximum at the point of inflection, and then starts declining, approaching zero as output approaches $X_2 = 0$. Therefore, given condition (6), if we choose a price line steeper than δ_1/δ_2 times the slope of the *PPF* at the inflection point, an internal production equilibrium will not exist. On the other hand, if we choose a price line flatter than δ_1/δ_2 times the slope of the *PPF* at the inflection point, two internal production equilibria will exist: one in convex to the origin range and the other in concave to the origin range of the *PPF*.

II. Implications of the Model

Using indifference curves to represent demand conditions, we can depict the overall equilibrium within a closed economy. It is defined by a tangency between the price line and an indifference curve at the production point and is represented by point B in Figure 1. Depending upon the actual specification of demand conditions and values of parameters δ_1 and δ_2, there may exist none, one, or more of such equilibria. Typically, one may expect two of them: one in convex to the origin range and the other in concave to the origin range of the *PPF*.

It has been shown in the paper by Jonathan Eaton and myself that a free trade equilibrium involving incomplete specialization and production in the convex to the origin range of the *PPF* is not always stable. For this reason, in what follows, attention will be focused mainly on the free trade equilibria involving either production in the concave to the origin range of the *PPF* or complete specialization in one commodity.

A. *A Small, Open Economy*

It has been noted by several authors that, if there are *IRS* in both industries, a small, open economy may specialize completely in one of them.[3] When there are *IRS* in one industry and *DRS* in the other, however, we have the following interesting result:

PROPOSITION 1: *Assuming homogeneous production functions, if there are IRS in one industry and DRS in the other, a small, open economy will never specialize completely in*

[2] Condition (6) is the same as that derived by Herberg and Kemp.

[3] See, for example, Wilfred Ethier; James Melvin; R. C. O. Matthews.

the IRS commodity. The economy may, however, specialize completely in the DRS commodity.

To prove this result, observe that given equations (1a) and (5) and $\delta_i \equiv 1/(1-\alpha_i)$, the firm's private costs of producing each good in terms of the other can be written as

$$\text{(7a)} \quad \frac{wL_{1j}}{p_2} = X_2^{(\delta_2-1)/\delta_2} X_1^{-(\delta_1-1)/\delta_1} X_{1j}$$

$$\text{(7b)} \quad \frac{wL_{2j}}{p_1} = X_1^{(\delta_1-1)/\delta_1} X_2^{-(\delta_2-1)/\delta_2} X_{2j}$$

From equations (7a) and (7b), we can obtain the firm's private marginal costs of production

$$\text{(8a)} \quad MC_{1j} = X_2^{(\delta_2-1)/\delta_2} X_1^{-(\delta_1-1)/\delta_1}$$

$$\text{(8b)} \quad MC_{2j} = X_1^{(\delta_1-1)/\delta_1} X_2^{-(\delta_2-1)/\delta_2}$$

Now as the economy approaches complete specialization in the *IRS* commodity, the output of the *DRS* commodity approaches zero. Remembering that $\delta_1 < 1$, equation (8b) yields

$$\text{(9)} \quad \lim_{X_1 \to 0} MC_{2j} = \infty$$

That is, as the economy approaches complete specialization in the *IRS* commodity, the private marginal cost of producing the latter approaches infinity. Therefore, unless the price of the *IRS* commodity is infinity, firms will not find it profitable to produce it near the point of complete specialization; for finite prices of the *IRS* commodity, specialization in it will necessarily be incomplete.

Next, note that complete specialization in the *DRS* commodity implies that the output of the *IRS* commodity be zero. Remembering that $\delta_2 > 1$, equation (8a) yields

$$\text{(10)} \quad \lim_{X_2 \to 0} MC_{1j} = 0$$

That is, as the economy approaches the point of complete specialization in the *DRS* commodity, the private marginal cost of producing the latter approaches zero. Therefore, so long as the price of the *DRS* commodity exceeds zero, firms will find it profitable to produce it near the complete specialization point. It follows that complete specialization in the *DRS* commodity cannot be ruled out.

It is instructive to illustrate Proposition 1 with the help of Figure 1. It was noted in Section I, Part A that the *PPF* exhibits increasing opportunity costs near the *IRS* axis. Moreover, as the production point moves closer to complete specialization in the *IRS* commodity, the slope of the *PPF* approaches zero. Therefore, it follows from equilibrium condition (6) that the economy will approach complete specialization in the *IRS* commodity only if the price of the *DRS* commodity approaches zero. In other words, for strictly positive commodity prices in the international market, the economy will never specialize completely in the *IRS* commodity. On the other hand, the marginal opportunity costs are declining near the *DRS* axis. Indeed, they approach zero as the economy approaches the point of complete specialization in the *DRS* commodity. Evidently, complete specialization in the *DRS* commodity is a possibility.

It is tempting to conjecture that if the extreme point of the *PPF* near the *DRS* axis happens to be sufficiently far out to the right, the economy might in fact benefit most by specializing completely in the *DRS* industry. Such a conjecture seems plausible at first, but turns out to be false on closer examination. While a rigorous proof of this assertion is provided in the Appendix, a simple geometric proof may be presented here.

Consider Figure 2 where line $P^0P^{0\prime}$ represents the international price ratio and P^0 is an internal production equilibrium. Corresponding to this price line and production point, we can construct a hypothetical *PPF* assuming as if the two industries were subject to constant returns to scale (*CRS*). Now, while the exact shape and position of the *PPF* will depend on the form of production functions that we specify, it will necessarily lie within or at most coincide with the price

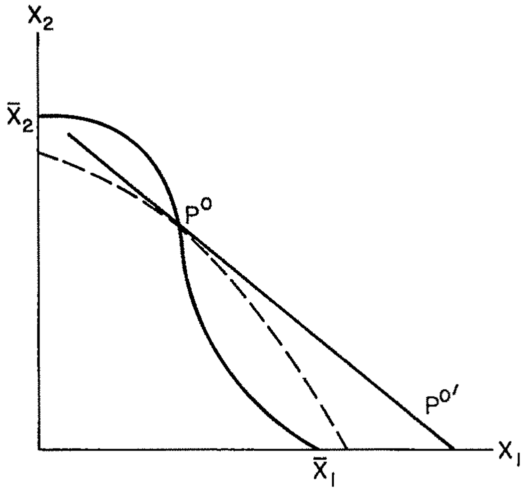


FIGURE 2

line $P^0P^{0\prime}$.[4] Evidently, when there are *DRS* in industry 1, the extreme point of the *PPF* near the X_1 axis must lie within the extreme point of the hypothetical *PPF* based on *CRS* in the two industries. This, of course, implies that the budget line passing through the point of complete specialization in the *DRS* industry must lie within $P^0P^{0\prime}$. Hence, the welfare level attained by complete specialization in the *DRS* industry will necessarily be inferior to that attained by incomplete specialization. We have

PROPOSITION 2: *If, at any given terms of trade, an internal production equilibrium exists, a welfare-maximizing small, open economy will never specialize completely in production.*

It must be noted that Proposition 2 is purely normative in nature. It only states which is the superior equilibrium from the welfare point of view but says nothing about which equilibrium will actually be reached. Since, under competition, individual consumers and producers—and not the country—are maximizers, it is possible for the country to be specialized completely in the *DRS* commodity after the opening of trade. In case of such an eventuality, the country will be consuming at a welfare level lower than what is potentially feasible.

It was noted in Section I, Part B that, at an internal production equilibrium, the private marginal product of a firm in the *DRS* industry exceeds the social marginal product in that industry while exactly the opposite is true in the *IRS* industry. Since competition leads to the equalization of private values of marginal products, it follows that, at an internal competitive equilibrium, the social value of marginal product in the *IRS* industry exceeds the social value of marginal product in the *DRS* industry. Therefore, an equilibrium such as P^0 in Figure 2, while superior to the equilibrium involving complete specialization in the *DRS* industry, is not optimum. Given higher social value of marginal product in the *IRS* industry, a tax subsidy scheme leading to an expansion of that industry and contraction of the *DRS* industry leads to improvement in welfare. This improvement in welfare is maximized when the tax subsidy scheme is just right to equalize the social values of marginal products in the two industries. Thus, we have

PROPOSITION 3: *Given output-generated IRS in one industry and DRS in the other, welfare maximization will require a permanent tax subsidy scheme encouraging expansion of the IRS industry and contraction of the DRS industry.*

Let us now turn to a comparison of the pretrade and posttrade equilibria. Does the opening of free trade result in an improvement in welfare? The answer to this question depends on where the pretrade and posttrade equilibria are located. If specialization in production is *incomplete* in the posttrade equilibrium and trade results in the expansion of the *IRS* industry, as shown by Kemp and Takashi Negishi in a multifactor, multicommodity model, the country unambiguously gains. If trade results in the expansion of the *DRS* industry, however, welfare implications are ambiguous.

[4] To be precise, given either one factor of production or two factors with identical factor intensities in the two industries, the *PPF* will coincide with the price line. If, given two factors of production, factor intensities in the two industries differ, however, the *PPF* will lie strictly inside the price line.

If, after the opening of trade, production is *completely* specialized in the *DRS* commodity, the welfare implications depend on the relative magnitudes of the pre- and posttrade price ratios. In Figure 1, if the posttrade relative price of the *DRS* commodity happens to be less than or equal to the autarkic price ratio $P^0P^{0'}$, the country will necessarily lose from trade. This is because budget lines through $\bar{X}_1$ having an absolute slope less than or equal to that of $P^0P^{0'}$ will necessarily lie inside the latter. The country will gain from trade if the relative price of the *DRS* commodity is sufficiently high to make the budget line through $\bar{X}_1$ steeper than $P^1P^{1'}$.

It may be useful at this point to discuss the implications of our analysis for the "Graham protection controversy." The main elements of this controversy have been elegantly summarized by Richard Caves, who characterized Graham as arguing that a country with a comparative advantage in the *DRS* industry and a comparative disadvantage in the *IRS* industry may be worse off under free trade than under autarky, or could at least improve its free trade welfare by protection. As Caves notes, Graham's arguments and numerical examples were insufficient to establish his case and he (Graham) met severe criticism from Frank Knight, among others. Years later, Tinbergen reformulated Graham's case and gave a geometric proof of it using the *PPF*. This reformulation seems to have settled the matter and, since then, Tinbergen's diagram has been reproduced at several places, some of which are cited in footnote 1.

It must be noted, however, that if production functions are assumed to be homothetic or homogeneous, Tinbergen's reformulation of Graham's case is incorrect. There are two major problems. First, he draws the *PPF* concave to the origin near the *DRS* axis and convex to the origin near the *IRS* axis. As a result, temporary protection leads to complete specialization in the *IRS* industry. Second, he does not recognize the presence of the divergence between the social and private marginal products at a competitive equilibrium.

The analysis in this paper provides the basis for correct reformulation of Graham's case for protection. It is clear that specialization in the *DRS* commodity as a result of the opening of trade may indeed result in a welfare loss; although it need not necessarily be so. It also follows from Proposition 2 that if, in the posttrade equilibrium, the country specializes completely in the *DRS* industry and an internal production equilibrium exists, it may be able to benefit by a temporary subsidy to the *IRS* industry. It must be noted, however, that while the temporary subsidy may *improve* welfare, it will not *maximize* the latter. As argued in Proposition 3, given output-generated economies and diseconomies of scale, maximization of welfare will require a permanent production-tax subsidy scheme.

Two final comments are in order. First, Graham had argued for protection through tariff. The present reformulation, on the other hand, argues for protection through a production-tax subsidy scheme. Second, Graham had argued that the country might gain from protection by *changing* its comparative advantage from the *DRS* industry to the *IRS* industry. The present analysis shows that a *change* in comparative advantage is not necessarily required for the gain; if the domestic demand for the *IRS* commodity is sufficiently strong, the country may continue to export the same (*DRS*) commodity after the subsidy as before it and still improve its welfare.

B. *A Two-Country Model*

I now discuss the implications of our production model for welfare and patterns of specialization in a two-country model. Suppose that the two countries *A* and *B* are characterized by identical tastes and technology. For simplicity, let the tastes be represented by the Mill-Graham utility function

$$(11) \qquad U_k = C_{1k}^{\beta} C_{2k}^{1-\beta} \qquad k = A, B$$

where C_{ik} denotes the consumption of commodity i in country k, β is a parameter such that $0<\beta<1$ and U_k is the utility level in country k. Finally, assume that country A is larger than B so that $L_A > L_B$. We then have

PROPOSITION 4: *Assuming that there are IRS in one industry and DRS in the other, and that tastes can be represented by the Mill-Graham utility function, in a two-country, two-commodity model with identical tastes and technology, free trade will result in the larger country exporting the IRS commodity and the smaller country exporting the DRS commodity. In the free trade equilibrium, the exporter of the IRS commodity will specialize incompletely while the exporter of the DRS commodity may specialize either completely or incompletely.*

In order to establish this pattern of trade and specialization, we must first derive the autarkic price ratios in the two countries. Denoting by P_k the equilibrium price of commodity 1 in terms of commodity 2 in country k, consumer equilibrium requires

$$(12) \qquad P_k = \frac{\partial U_k/\partial C_{1k}}{\partial U_k/\partial C_{2k}} = \frac{\beta}{1-\beta}\frac{C_{2k}}{C_{1k}} \qquad k=A,B$$

Moreover, from equation (5), equilibrium on the production side of the economy yields

$$(13) \qquad P_k = \frac{X_2^{(\delta_2-1)/\delta_2}}{X_1^{(\delta_1-1)/\delta_1}}$$

In the absence of trade, we must have $C_{ik} = X_{ik}$. Therefore, equations (1b) and (2) (with proper country subscript on variables) can be combined with equations (12) and (13) to obtain[5]

$$(14) \qquad \frac{L_{1k}}{L_{2k}} = \frac{\beta}{1-\beta}$$

$$(15) \qquad P_k = \frac{\beta^{1-\delta_1}}{(1-\beta)^{1-\delta_2}} L_k^{\delta_2-\delta_1}$$

Setting subscript k equal to A and B in equation (15) and dividing, we have

$$(16) \qquad \frac{P_A}{P_B} = (L_A/L_B)^{\delta_2-\delta_1}$$

Remembering that $L_A > L_B$ and $\delta_2 > \delta_1$, it follows from equation (16) that $P_A > P_B$. Thus, in autarky, the *DRS* commodity is more expensive in the larger country. Therefore, when trade opens it will export the *IRS* commodity and import the *DRS* commodity. The smaller country will, in turn, export the *DRS* commodity and import the *IRS* commodity.

The intuitive explanation of this result lies in the fact that given identical Mill-Graham utility functions, the two countries allocate labor between industries 1 and 2 in the same proportion. This means that the larger country produces more of both the commodities. Remembering that the larger the scale of production, the higher (lower) the marginal cost of producing the *DRS* (*IRS*) commodity, it is evident that, in autarky, the *DRS* (*IRS*) commodity will be more (less) expensive in the larger country.

It was shown in Section II, Part A that as a country approaches the point of complete specialization in the *IRS* commodity, the private marginal cost of producing the latter approaches infinity. For this reason, the exporter of the *IRS* commodity, country A, will remain incompletely specialized in the posttrade equilibrium.

As regards country B, it may specialize either completely or incompletely in production. Other things being equal, the more disparate the two countries in terms of factor endowments, the greater the likelihood that B will specialize completely. This is because the more disparate the two countries, from equation (15), the larger will be

[5] Note that given the Mill-Graham utility function we have a unique autarkic equilibrium. This equilibrium may lie in either concave or convex range of the *PPF*. More precisely, observe that setting the right-hand side of equation (4) equal to zero, the inflection point on country k's *PPF* is given by $L_{1k}/L_{2k} = (\delta_1 - 1)/(1-\delta_2)$. Comparing this with equation (14), it can be deduced that the equilibrium will lie in the concave or convex to the origin range of the *PPF* as $\beta/(1-\beta) \lesseqgtr -(\delta_1-1)/(\delta_2-1)$.

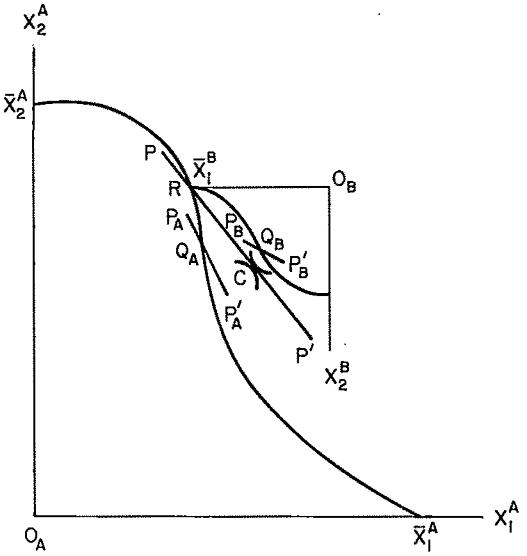


FIGURE 3

the difference between their autarkic price ratios. Moreover, the larger is country A relative to country B, the closer will the posttrade price ratio be to the former's pretrade price ratio. Thus, the greater disparity of size will lead to a larger increase in the price of the DRS commodity in country B after the opening of trade. Since, beyond the inflection point on the PPF, the private marginal costs of production are decreasing, a large enough increase in the price of the DRS commodity may induce the firms to specialize completely in the DRS commodity. Indeed, if country A is sufficiently large relative to country B, the posttrade price of the DRS commodity may be high enough to rule out the existence of an internal equilibrium in the latter.[6] In case of such eventuality, specialization in the DRS commodity will necessarily be complete.

The possibility of complete specialization by the exporter of the DRS commodity is depicted in Figure 3 where the PPF of country A has been drawn in the normal fashion while that of country B has been turned upside down and placed with its X_1 axis parallel to the corresponding axis of A and with point $\bar{X}_1^B$ on the top of point R of A's PPF. In autarky, production and consumption in countries A and B take place at points Q_A and Q_B, respectively. The pretrade price ratios in the two countries are given by $P_AP'_A$ and $P_BP'_B$, respectively. Trade results in the international price ratio PP'. At this price ratio, no internal production equilibrium exists in country B. Therefore, it specializes completely in the DRS commodity. On the other hand, country A specializes incompletely in the IRS commodity and produces at point R. Both countries consume at point C. Note that country A necessarily gains from trade. Since it exports the IRS commodity, that industry must expand and the DRS industry must contract. With incomplete specialization, this entails an improvement in welfare.[7] The effect of trade on the welfare of country B is ambiguous in general, but as drawn in Figure 3 it does benefit. Observe that before trade opens, compared to the Pareto optimal production point, B is underproducing the IRS commodity and overproducing the DRS commodity. Free trade, by contracting the former and expanding the latter, drives the economy further away from the Pareto optimal allocation. This results in a loss of welfare to the economy. But at the same time, the economy gains from trade; the price of the (DRS) commodity it exports is higher in the free trade equilibrium than under autarky. On balance, the country will gain or lose accordingly as the additional gain from trade offsets or is offset by the additional loss due to the expansion of the DRS industry and contraction of the IRS industry.

III. The Multifactor Case

The first three propositions of this paper remain valid in the presence of more than

[6] It was noted earlier that the maximum price an internal equilibrium can support is given by δ_1/δ_2 times the slope of the PPF at the inflection point. Furthermore, it can be shown that the smaller the labor endowment, the lower is the maximum price. Therefore, given that country B is sufficiently small relative to country A, even though at the international price ratio an internal equilibrium exists for the latter, it may not exist for the former.

[7] This follows from a theorem by Kemp and Nagishi referred to earlier in Section II, Part A.

one factor of production. This follows from the fact that all the properties of the *PPF* and the relationship between its slope and the price line continue to hold in the multifactor case. It has been shown by Herberg and Kemp that even when we allow for n factors of production, the *PPF* remains strictly convex to the origin near the *DRS* axis and strictly concave to the origin near the *IRS* axis.[8] Moreover, it is easily verified that irrespective of the number of factors, given homogeneous production functions, the marginal cost of producing the *IRS* good approaches infinity in the neighborhood of $X_1=0$ and that of the *DRS* good approaches zero in the neighborhood of $X_2=0$.

As regards Proposition 4, it has to be modified slightly in the multifactor case to take account of the possibility of multiple equilibria in the pretrade situation. The patterns of trade in this case depend on which autarkic equilibria are chosen for the comparison of pretrade prices in the two countries. As far as the possibility of complete specialization by the exporter of the *DRS* commodity and incomplete specialization by the exporter of the *IRS* commodity is concerned, it remains valid in the multifactor case.

The main difference between one-factor and multifactor models is that in the latter case we may have more than one point of inflection in the intermediate range of the *PPF.* But this possibility in the multifactor case does no harm to our results; for they do not *require* a unique point of inflection in the *PPF.* The only significance of this property is to make the diagrammatic analysis more elegant.

IV. Summary of Results

This paper has analyzed in detail the production model based on *IRS* in one industry and *DRS* in the other. Assuming homogeneous production functions the following propositions have been made. First, if a small, open economy specializes completely in production, it will do so in the *DRS* commodity. Second, at any given terms of trade, if an internal production equilibrium exists, a welfare-maximizing small, open economy will never specialize completely in production. Third, welfare maximization will require a permanent tax subsidy scheme. Finally, in a two-country model, international trade may result in complete specialization by the exporter of the *DRS* commodity and incomplete specialization by the exporter of the *IRS* commodity.

Appendix

This Appendix proves that the point of complete specialization in the *DRS* commodity, $\overline{X}_1$ in Figure 2, lies inside the price line ($P^0P^{0\prime}$) that goes through a feasible point of incomplete specialization. Denoting the slope of the price line by $-P(\equiv -p_1/p_2)$ and the coordinates of the equilibrium point P^0 in Figure 2 by (X_1^0, X_2^0), the equation of $P^0P^{0\prime}$ may be written as

$$(X_2-X_2^0)=-P(X_1-X_1^0) \tag{A1}$$

where (X_1,X_2) is any point on the line. Setting $X_2=0$ in (A1), we find that $P^0P^{0\prime}$ intersects the X_1 axis at

$$X_1=\frac{1}{P}X_2^0+X_1^0 \tag{A2}$$

Taking account of equations (1)–(3) and (5), (A2) reduces to

$$X_1=L_1^{0^{\delta 1-1}}L \tag{A3}$$

where L_1^0 is the amount of labor employed in industry 1 at point (X_1^0, X_2^0).

When the economy is completely specialized in industry 1, the output is given by

$$\overline{X}_1=L^{\delta_1} \tag{A4}$$

comparing (A3) and (A4), it is easily verified that $\overline{X}_1<X_1$.

REFERENCES

Richard E. Caves, *Trade and Economic Structure: Models and Methods,* Cambridge, Mass. 1960, 169–74.

[8] See Herberg and Kemp (Theorem 2′, p. 414).

______ **and Ronald W. Jones,** *World Trade and Payments*, Boston 1973.

J. Eaton and A. Panagariya, "Gains from Trade under Variable Returns to Scale, Commodity Taxation, Tariffs and Factor Market Distortions," *J. Int. Econ.*, Nov. 1979, *9*, 481–501.

W. Ethier, "Increasing Returns to Scale and International Trade," mimeo., Univ. Pennsylvania 1978.

F. Graham, "Some Aspects of Protection Further Considered," *Quart. J. Econ.*, Feb. 1923, *37*, 199–227.

H. Herberg and M. C. Kemp, "Some Implications of Variable Returns to Scale," *Can. J. Econ.*, Aug. 1969, *2*, 403–15.

M. C. Kemp and T. Negishi, "Variable Returns to Scale, Commodity Taxes, Factor Market Distortions and Their Implications for Gains from Trade," *Swedish J. Econ.*, Jan. 1970, *72*, 1–11.

Charles P. Kindleberger, *International Economics*, 4th ed., Homewood 1968.

F. H. Knight, "Some Fallacies in the Interpretation of Social Cost," *Quart. J. Econ.*, Aug. 1924, *38*, 592–606.

R. C. O. Matthews, "Reciprocal Demand and Increasing Returns," *Rev. Econ. Stud.*, Feb. 1950, *17*, 149–58.

J. R. Melvin, "Increasing Returns to Scale as Determinants of Trade," *Can. J. Econ.*, Aug. 1969, *2*, 389–402.

C. E. Staley, *International Economics*, Englewood Cliffs 1970.

Jan Tinbergen, *International Economic Cooperation*, Amsterdam 1945.

Ingo Walter, *International Economics*, New York 1975.

Dynamic Models of Portfolio Behavior: A General Integrated Model Incorporating Sequencing Effects

By P. DORIAN OWEN*

In a recent article in this *Review*, Douglas Purvis advocates an "integrated" model of savings and portfolio allocation as an extension of the "pitfalls" model developed by William Brainard and James Tobin. Brainard and Tobin (hereafter, B-T) propose a multivariate stock-adjustment model in which, generally, changes in the holdings of any one asset depend not only on its own deviation of actual value from desired or target value but also on the corresponding deviations for other assets. They also argue that attention should be paid to balance sheet or income identities when specifying and estimating the equation system, hence generating "consistency" or "adding-up" requirements. They take the total level of end-of-period wealth for any sector as predetermined, given by the total of asset holdings inherited from the previous period plus savings and capital gains which accrue in the current period. By taking the change in wealth for any sector as predetermined, B-T separate the consumption-savings decision from the portfolio-allocation decision. Purvis argues that such separation is overrestrictive and, in the presence of adjustment costs, likely to be invalid. He therefore proposes an integrated model of savings and portfolio allocation embodying the key features of the B-T approach but with a wider relevant budget constraint. A key feature of this integration is that portfolio-composition effects enter the consumption-savings decision.

However, Gary Smith (1978) points out that there is a separate issue from that of the existence of portfolio-composition effects on consumption, namely the "sequencing" or hierarchy of decision. Sequencing depends on variables influencing the consumption-savings decision which are not directly relevant to portfolio-allocation decisions, although they may have indirect effects through their influence on total wealth. Purvis' model neglects the possibility of such influences. Furthermore, Smith argues that Purvis' extension, by implicitly attributing a "simple" consumption function (which excludes lagged asset holdings) to sequential decision making, misrepresents B-T.[1] Hence, all comparisons that Purvis makes with B-T must be conditional on the assumed form of the B-T consumption function. In his discussion of the relationship between the integrated and sequential approaches, Smith attempts to show that Purvis' model, despite advocating a more general budget constraint and explicitly considering the consumption-savings decision, is a special case of the B-T model.

The aim of this paper is to reconcile as far as possible the approaches of Purvis and Brainard and Tobin in the context of a general model embodying the most relevant features of both. In Section I, a more general model is developed which jointly determines the consumption-savings decision and the portfolio-allocation decision subject to an income constraint. The model allows for both sequencing effects and portfolio-composition effects on consumption. An appropriate set of adding-up restrictions is derived and both the B-T and Purvis models are seen to be special cases of the general

*University of Reading. I am grateful to Douglas Purvis, Gary Smith, and the referee for helpful comments on earlier drafts of this paper.

[1] "The substance of ...[Brainard and Tobin's] ...sequential approach is not that the composition of wealth is unimportant to consumption but rather that there are some variables which influence consumption and yet do not separately affect asset demands; only the net amount of saving motivated by these influences is important" (see Smith, 1978, p. 410).

model. In particular, my model provides an alternative interpretation of the B-T hierarchical approach as derived from a general integrated approach with certain, not unreasonable, a priori parameter restrictions. In Section II Smith's observations on the level of generality of Purvis' integrated model are reexamined in the context of the more general model and are shown to be misleading. While both the B-T and Purvis models are special cases of my general model, neither is a special case of the other. Section III includes comments on some of Purvis' general conclusions and on estimation problems.

I. A General Model Incorporating Sequencing and Portfolio-Composition Effects on Consumption

Consider a specific sector of the economy which can allocate its wealth between n different assets and liabilities (y_r, $r=1, \ldots, n$). Let x be the row vector of explanatory variables relevant either to the consumption-savings decision and/or the portfolio-allocation decision. Without loss of generality we can partition x into $(x_1|x_2)$; where x_1 includes the $(m+1)$ explanatory variables $X_0, X_1, \ldots, X_m$, that are assumed to be relevant in the asset demand equations. $X_m(=\Sigma_{s=1}^n y_s)$ is total wealth at the end of the period, X_0 is income, and X_i, $i=1,\ldots,m-1$, would include, for example, interest rates on the various assets and any other relevant variables. Some or all of these variables may also be relevant to the consumption-savings decision. The vector x_2 includes the k explanatory variables $X_{m+1},\ldots,X_{m+k}$ which, *if they exist*, are relevant to the consumption-savings decision, but are not separately relevant to portfolio-allocation decisions—they influence asset holdings only through wealth.[2] Initially, however, we shall consider X_i, $i=1,\ldots, m-1, m+1,\ldots,m+k$, as potentially relevant to both the asset-demand equations and the consumption function. We can then impose a priori restrictions when considering the B-T model as a special case.

Brainard and Tobin, Purvis, and Smith consider end-of-period equilibrium models involving unobservable long-run "desired" or target levels of asset holdings which are assumed to depend on a set of observable explanatory variables. They propose a multivariate stock-adjustment process:

$$(1)\qquad \Delta y_r = \sum_{s=1}^{n} \gamma_{rs}[y_s^* - y_s(-1)] \qquad r=1,\ldots,n$$

where $y_r(-1)$ and y_r are beginning- and end-of-period holdings of the rth asset, respectively, and y_r^* is the desired holding of the rth asset. Following Brainard and Tobin, let us assume that[3]

$$(2)\qquad y_s^* = \sum_{i=0}^{m+k} \beta_{si}X_i \qquad s=1,\ldots,n$$

Substituting (2) in (1) and adding a stochastic disturbance term:[4, 5]

$$(3)\quad \Delta y_r = \sum_{i=0}^{m+k} \alpha_{ri}X_i - \sum_{s=1}^{n} \gamma_{rs}y_s(-1)+\varepsilon_r \qquad r=1,\ldots,n$$

[2] The notation is that used by Purvis and Smith (1978). The only additions are the variables $X_{m+1}, \ldots, X_{m+k}$. Smith considers the existence of such variables but does not explicitly include them in the model.

[3] Smith (1975) and B-T consider a form of (2) where desired holdings of the various assets and liabilities are homogeneous of the first degree in wealth. This is not critical to the following arguments.

[4] Brainard and Tobin also include ΔX_m in their equivalent of (1). Smith (1975) shows that ΔX_m is redundant (unless, as in B-T, one of the asset discrepancies is eliminated). Brainard and Tobin assume that desired holdings of assets sum to total wealth, hence there is an exact linear relationship between $X_m = \Sigma_s y_s^*$, $X_m(-1) = \Sigma_s y_s(-1)$, and ΔX_m. As Purvis correctly points out $\Sigma_r y_r^* = X_m$ is not a budget constraint since it does not necessarily hold, rather it is an additional behavioral assumption (a rational desires hypothesis). This assumption is not a necessary condition for the redundancy of ΔX_m in my model; the identity $X_m - X_m(-1) \equiv \Delta X_m$ is sufficient for this purpose.

[5] Stochastic disturbance terms are not explicitly included in (3) by B-T or Smith (1975) nor in Purvis' equations. They are introduced into the argument by Smith (1978).

where $a_{ri}=\sum_{s=1}^{n} g_{rs}b_{si} \qquad i=0,\dots,m+k$

For the consumption function consider:[6]

$$(4) \qquad C=\sum_{i\neq m} b_iX_i+\sum_{s=1}^{n} e_sy_s(-1)+\varepsilon_0$$

where C is consumption.

It is explicitly assumed that $b_m=0$. End-of-period wealth is a result of the consumption-savings decision not a determinant of it. The influence of beginning-of-period wealth is incorporated in $y_s(-1)$; $s=1,\dots,n$. It is possible that some of the b_i, $i\neq m$, are zero, but this is not imposed a priori. Note that (3) includes X_m for the reasons suggested by Smith. The relevant income constraint is taken to be[7]

$$(5) \qquad X_0=C+\sum_{r=1}^{n}\Delta y_r$$

Next we derive the adding-up constraints for this model. Substitute (3) and (4) in (5):

$$(6) \quad X_0=\sum_{i\neq m} b_iX_i+\sum_{s=1}^{n} e_sy_s(-1)+\varepsilon_0$$
$$+\sum_{r=1}^{n}\sum_{i=0}^{m+k}\alpha_{ri}X_i-\sum_{r=1}^{n}\sum_{s=1}^{n}\gamma_{rs}y_s(-1)+\sum_{r=1}^{n}\varepsilon_r$$

An implication of the model is that changing X_i, $i\neq m$, $y_s(-1)$, $s=1,\dots,n$ or ε_0 in (6) will affect C (unless the relevant parameter happens to be zero) and therefore affects X_m. For example, consider increasing X_0 by one unit. This implies that C increases by b_0 units and savings, and hence X_m, increase by $(1-b_0)$ units. We assume $0<b_0<1$. Similarly, increasing $X_i(i\neq 0,m)$ by one unit implies that C increases by b_i units and savings and X_m decrease by b_i units. Therefore, for any variable X_i $(i\neq m)$ which is relevant to the consumption-savings decision and also relevant to the portfolio-allocation decision, any change in X_i will have an "indirect" effect on Δy_r through a change in the allocation of the given level of income[8] between consumption and savings (channelled through the X_m term in the asset-demand equation), as well as a "direct" effect.[9] Taking into account both these effects, (6) always holds if and only if:[10]

$$(7a) \quad \sum_{r=1}^{n}\alpha_{r0}+b_0\left(1-\sum_{r=1}^{n}\alpha_{rm}\right)+\sum_{r=1}^{n}\alpha_{rm}=1$$

$$(7b) \quad \sum_{r=1}^{n}\alpha_{ri}+b_i\left(1-\sum_{r=1}^{n}\alpha_{rm}\right)=0 \quad i\neq 0,m$$

$$(7c) \quad \sum_{r=1}^{n}\gamma_{rs}-e_s\left(1-\sum_{r=1}^{n}\alpha_{rm}\right)-\sum_{r=1}^{n}\alpha_{rm}=0 \qquad s=1,\dots,n$$

$$(7d) \quad \sum_{r=1}^{n}\varepsilon_r+\varepsilon_0\left(1-\sum_{r=1}^{n}\alpha_{rm}\right)=0$$

In terms of the above model Purvis makes the a priori assumption that $\alpha_{rm}=0$ for all r (therefore $\Sigma_{r=1}^{n}\alpha_{rm}=0$).[11] He states, "...by

[6] Throughout, $i\neq m$ and $i\neq 0$, m correspond to $i=0$, $1,\dots,m-1$, $m+1,\dots,m+k$; and $i=1,\dots,m-1$, $m+1,\dots,m+k$, respectively.

[7] The model could be made more general by making explicit the role of capital gains and losses, both in the budget constraint and in the asset equations. However, this would not affect the following arguments and I have chosen to preserve the basic framework adopted by Purvis and Smith except where critical to the argument. For the household sector, the widest possible budget constraint would also have labor income endogenous.

[8] For $i=0$, increases in income will lead to an increase in C and an increase in $X_m(0<b_0<1)$.

[9] Brainard and Tobin consider only the direct effects. This shows the danger of deriving adding-up restrictions in a sequential model where the consumption-savings function is not made explicit. Ironically, this can be compared to one of B-T's important conclusions, namely that the specification of any residual equations should be directly checked—they, of course, were referring to residual asset equations within the portfolio, but the same warning also applies to the consumption function.

[10] Note that there is no adding-up restriction containing the coefficients of the wealth variable *only*. *In* the integrated model under consideration, X_m can change only if a variable relevant to the consumption-savings decision changes (i.e., X_i; $i\neq m$; $y_s(-1)$; $s=1,\dots,n$ or ε_0).

[11] Strictly, Purvis includes $X_0,\dots,X_{m-1}$ in his asset equations. However, the flavor of Purvis' approach is

assumption current wealth (X_m) does *not* enter the long-run asset demand functions" on the grounds that "...it is the composition of wealth which is important in determining the optimum time paths" (p. 405). Substituting this assumption into (7) gives Purvis' adding-up restrictions as a special case:

$$(8a) \quad b_0 + \sum_{r=1}^{n} \alpha_{r0} = 1$$

$$(8b) \quad b_i + \sum_{r=1}^{n} \alpha_{ri} = 0 \qquad i \neq 0, m$$

$$(8c) \quad e_s - \sum_{r=1}^{n} \gamma_{rs} = 0 \qquad s = 1, \ldots, n$$

$$(8d) \quad \varepsilon_0 + \sum_{r=1}^{n} \varepsilon_r = 0$$

Next consider Smith's version of the B-T model. The consumption function is never made explicit, but it could not be a more general linear function than (4). A key feature of Smith's arguments is the existence of, variables relevant to the consumption-savings decision but not relevant to the portfolio-allocation decision other than through the level of wealth generated. This brings in the distinction between x_1 and x_2 and corresponds to the a priori assumption that $\alpha_{ri} = 0$, $i = m+1, \ldots, m+k$, for all r (which implies $\Sigma_{r=1}^{n} \alpha_{ri} = 0$, $i = m+1, \ldots, m+k$).[12] Therefore, for $i = m+1, \ldots, m+k$, (7b) becomes

$$(9) \qquad b_i \left(1 - \sum_{r=1}^{n} \alpha_{rm}\right) = 0$$

From (9) either $b_i = 0$ and/or $\Sigma_{r=1}^{n} \alpha_{rm} = 1$. Hence, *if* there exists *any* explanatory variable which is relevant to the consumption-savings decision, *and if* this variable is not also relevant in the portfolio-allocation decision (i.e., it does not appear as an explanatory variable in the equation for Δy_r *for any* r), then $\Sigma_{r=1}^{n} \alpha_{rm} = 1$ is a required adding-up restriction. It is difficult to say a priori whether both these conditions will be met by any variable. Smith cites a number of candidates such as the composition of income, commodity prices, and accustomed consumption levels. Certainly it seems reasonable that these are potentially important to the consumption-savings decision. The key question, however, is whether or not they are separately relevant in the equation for at least one of the Δy_r.[13] In this respect, omission of relevant explanatory variables, in order to keep the number of variables within manageable limits (hence incurring possible misspecification bias on estimation), is quite different from omission of variables because they are not relevant.[14]

However, if we accept that $\Sigma_{r=1}^{n} \alpha_{rm} = 1$, (7) simplifies to

$$(10a) \quad \sum_{r=1}^{n} \alpha_{ri} = 0 \qquad i = 0, 1, \ldots, m-1$$

$$(10b) \quad \sum_{r=1}^{n} \alpha_{rm} = 1$$

$$(10c) \quad \sum_{r=1}^{n} \gamma_{rs} = 1 \qquad s = 1, \ldots, n$$

$$(10d) \quad \sum_{r=1}^{n} \varepsilon_r = 0$$

that he makes no distinction between x_1 and x_2 and, therefore, we have expanded the set of variables in his model to include all the exogenous variables which affect consumption and portfolio decisions.

[12] The following results also hold in the case where $\alpha_{ri} = 0$ may not be true for all r, $i = m+1, \ldots, m+k$, but $\Sigma_{r=1}^{n} \alpha_{ri} = 0$, $i = m+1, \ldots, m+k$.

[13] It is not unreasonable to argue that such factors might be relevant in some asset-demand functions: for example, considering the personal sector, accustomed consumption levels may have an influence on desired holdings of liquid assets relative to illiquid assets; the composition of income could influence allocation of any savings, particularly if a concept of income that includes capital gains is adopted etc.

[14] We are, of course, considering the values of the unknown population parameters, not estimated values. Presumably, following part of the usual justification for including stochastic disturbance terms, the influence of a large number of variables relevant but not critically important to C and Δy_r, $r = 1, \ldots, n$, are included in the ε_0 and ε_r terms.

which are the B-T adding-up restrictions. The coefficients b_i, $i \neq m$, and e_s, $s=1,\ldots,n$, drop out and are not constrained. Therefore, we obtain the B-T adding-up restrictions when there exists an explanatory variable X_i such that $\alpha_{ri}=0$, for all r, and $b_i \neq 0$. This is consistent with Smith's "...simplifying assumption that many consumption influences are not separately important to portfolio decisions" (1978, p. 415). It is important to note, however, that this result has been derived from a framework in which the consumption function is made explicit and the derivation of the adding-up restrictions has taken into account the interaction between the consumption-savings decision and portfolio-allocation decision. It thus enables us to view the B-T and Purvis models within a common general framework. Furthermore, even for the case where the adding-up restrictions in (7) simplify to the B-T restrictions, we are not assuming a predetermined level of wealth. The restrictions in (10) arise as a result of making certain a priori assumptions about some coefficients in the general integrated model in which consumption, savings, and wealth are fully explained, assuming a predetermined level of income.[15]

The B-T restrictions can also be derived from our general model if ε_0 can change, with the ε_r's held constant, in the mental *ceteris paribus* experiments underlying the derivation of (7). If this is possible, (7d) implies

$$\varepsilon_0\left(1-\sum_{r=1}^{n}\alpha_{rm}\right)=0$$

and, if $\varepsilon_0 \neq 0$, $\Sigma_{r=1}^{n}\alpha_{rm}=1$. Furthermore, the B-T restrictions can be obtained if ε_0 affects individual ε_r's but not their sum, $\Sigma_{r=1}^{n}\varepsilon_r=0$. If, for instance, $\varepsilon_r=g_r\varepsilon_0+u_r$, $r=1,\ldots,n$, with $\Sigma_{r=1}^{n}g_r=\Sigma_{r=1}^{n}u_r=0$, then X_m could change with ε_r constant and the B-T restrictions would then be necessary and sufficient. It is difficult to evaluate a priori whether either of these situations is likely to occur in reality because of the unobservable nature of the stochastic disturbance terms. In the absence of a theoretical explanation of the nature of the ε terms in this model, it would appear that a set of general constraints like (7) which allow changes in ε_0 to affect Δy_r through changes in X_m *and* changes in ε_r is, initially, more appropriate. Elimination of either of these routes of influence could be considered as yielding a special case of our model.[16]

The B-T restrictions could very well be valid in reality. However, *in principle*, it would be desirable to estimate the more general model subject to (7), particularly if we are interested in the interaction between the consumption function and the asset-demand equations. We could then test any exclusion restrictions.

II. Model Generality

The purpose of this section is to reconsider some of Smith's criticisms of Purvis' model in terms of our more general model. Many of the points raised by Smith have already been taken into account in formulating the model, particularly the inclusion of X_m in the equation for Δy_r to allow for the effect of ε_0 and any variables relevant only to the consumption-savings decision.

However, some interesting results occur if the procedure adopted by Smith to show the greater generality of the B-T model is applied to our more general integrated model. It is possible to express our model in a sequential form with asset demands depending upon wealth only, rather than wealth and income. Substituting (4) into (5) and

[15] Compare Smith's comment on the B-T equations: "The model and the restrictions in fact contain no information about the determination of wealth" (1978, p. 413, fn. 3).

[16] I am grateful to Gary Smith for pointing out these two additional sets of conditions under which the B-T restrictions can be derived from our general model. I am, of course, responsible for any errors in my interpretation of his comments. As yet there has been little detailed work on the source of disturbances and the way their effects enter equation systems with singular variance-covariance disturbance matrices. Developments in this direction could have important implications for our interpretation of the sort of *ceteris paribus* experiments carried out here, as well as obvious advantages for estimation.

assuming $(1-b_0)\neq 0$:

$$X_0 = \left(X_m + \sum_{i=1}^{m-1} b_i X_i + \sum_{i=m+1}^{m+k} b_i X_i + \sum_{s=1}^{n} (e_s - 1) y_s(-1) + \varepsilon_0 \right) \Big/ (1-b_0)$$

Substituting for X_0 in (3):

$$\Delta y_r = X_m \left(\frac{\alpha_{r0}}{1-b_0} + \alpha_{rm} \right) + \sum_{i\neq 0,m} \left(\alpha_{ri} + \frac{\alpha_{r0} b_i}{1-b_0} \right) X_i - \sum_{s=1}^{n} \left(\gamma_{rs} + \frac{\alpha_{r0}}{1-b_0}(1-e_s) \right) y_s(-1) + \left(\varepsilon_r + \frac{\alpha_{r0}}{1-b_0} \varepsilon_0 \right)$$

which can be written in the form:

$$(11)\quad \Delta y_r = \sum_{i=1}^{m+k} \bar{\alpha}_{ri} X_i - \sum_{s=1}^{n} \bar{\gamma}_{rs} y_s(-1) + \bar{\varepsilon}_r$$

where $\bar{a}_{ri} = a_{ri} + \dfrac{a_{r0} b_i}{1-b_0}$ $\quad i\neq 0, m$, etc.

The definitions of the parameters $\bar{\alpha}_{ri}$ etc. in (11) and the adding-up restrictions in (7) imply the adding-up restrictions for the sequential version of the general integrated model:

$$(12a)\quad \sum_{r=1}^{n} \bar{\alpha}_{ri} = 0 \qquad i\neq 0, m$$

$$(12b)\quad \sum_{r=1}^{n} \bar{\alpha}_{rm} = 1$$

$$(12c)\quad \sum_{r=1}^{n} \bar{\gamma}_{rs} = 1 \qquad s=1,\ldots,n$$

$$(12d)\quad \sum_{r=1}^{n} \bar{\varepsilon}_r = 0$$

Initially, it appears that Smith's arguments are valid not only for Purvis' model, but also for the more general integrated model. The sequential version of the latter appears to be the special case of the B-T model in which $\bar{\alpha}_{r0}=0$! (Compare (12) with (10).) An even more surprising result can be obtained if the general integrated model is expressed in terms of Purvis' framework where X_m is not included in the asset-demand equations. Again substituting (4) in (5):

$$(13)\quad X_m = X_0 - C + \sum_{s=1}^{n} y_s(-1) = X_0(1-b_0) - \sum_{i\neq 0,m} b_i X_i - \sum_{s=1}^{n} (e_s - 1) y_s(-1) - \varepsilon_0$$

Equation (13) can be used to substitute for X_m in the asset-demand equations (3):

$$(14)\quad \Delta y_r = (\alpha_{r0} + \alpha_{rm}(1-b_0)) X_0 + \sum_{i\neq 0,m} (\alpha_{ri} - \alpha_{rm} b_i) X_i - \sum_{s=1}^{n} (\gamma_{rs} + \alpha_{rm}(e_s - 1)) y_s(-1) + (\varepsilon_r - \alpha_{rm}\varepsilon_0)$$

which can be written in the form:

$$(15)\quad \Delta y_r = \sum_{i\neq m} \alpha_{ri}^{**} X_i - \sum_{s=1}^{n} \gamma_{rs}^{**} y_s(-1) + \varepsilon_r^{**}$$

where $\alpha_{ri}^{**} = (a_{ri} - a_{rm} b_i)$, $\quad i\neq 0$, m, etc.

The definitions of the parameters α_{ri}^{**} etc. in (15) and the adding-up restrictions in (7) imply the adding-up restrictions for the form of the general integrated model which is similar to Purvis' model in that X_m has been substituted out:

$$\text{(16a)} \quad b_0 + \sum_{r=1}^{n} \alpha_{r0}^{**} = 1$$

$$\text{(16b)} \quad b_i + \sum_{r=1}^{n} \alpha_{ri}^{**} = 0 \qquad i \neq 0, m$$

$$\text{(16c)} \quad e_s - \sum_{r=1}^{n} \gamma_{rs}^{**} = 0 \qquad s = 1, \ldots, n$$

$$\text{(16d)} \quad \varepsilon_0 + \sum_{r=1}^{n} \varepsilon_r^{**} = 0$$

Following Smith's argument this seems to imply that the model in (15) and (4), which is equivalent to the model in (3) and (4), is of comparable generality with Purvis' model. (Compare (16) with (8).) However, as shown in Section I, Purvis' model is the special case of the model in (3) and (4) with $\alpha_{rm} = 0$, for all r. Hence there is an obvious contradiction. The income constraint (5) ensures an exact linear relationship between the predetermined variables X_0 and $\Sigma_{r=1}^{n} y_r(-1)$ and the *ex post* values of C and X_m, hence allowing any one of these variables to be substituted by the other three in the asset-demand equations. However, the method used by Smith to show the allegedly greater generality of the B-T model is not reliable. A more accurate indication of whether a particular model is more or less general than some other model is given by observing what a priori restrictions are made on the coefficients in the general integrated model in Section I. In this respect both the B-T and Purvis models are less general than our model. However, they are nonnested and, in terms of our model, it is not possible to say, a priori, which is more general without views on which makes the more "extreme" assumptions.

III. Other General Comments

The model outlined in Section I provides an appropriate starting point for empirical analysis of a consumption function-asset equations system and a framework in which any a priori restrictions can be categorized and, in principle, tested. Such a model is preferable to Purvis' model because it explicitly accounts for the possibility that there may exist variables which are relevant to the consumption-savings decision, but are relevant to the portfolio-allocation decision only in their influence on the level of wealth to be allocated. Nevertheless, many of Purvis' general conclusions are still valid. For example, Smith argues that "...consistency in a pure linear own-adjustment model requires that actual holdings adjust fully so as to always coincide with desired holdings" (1975, p. 511). The key reason for this is the requirement that $\Sigma_{r=1}^{n} y_r^* = X_m$ holds. Because this extra behavioral assumption is not required in our model, univariate adjustment (where $\gamma_{rs} = 0, r \neq s$; $\gamma_{rr} = e_r + \Sigma_{r=1}^{n} \alpha_{rm}(1 - e_s)$) and a "constant" speed of adjustment (where $e_s = \bar{e}$ for all s, $\Sigma_{r=1}^{n} \gamma_{rs} = \bar{e} + \Sigma_{r=1}^{n} \alpha_{rm}(1 - \bar{e})$) are both possible special cases. The adoption of either of these special cases would have to be justified on empirical grounds; they could not be ruled out on a priori theoretical grounds.

The B-T approach has been adopted in many empirical studies of specific sectors of the economy (for example, see Benjamin Friedman; R. W. Kopcke; Mitsuo Saito; W. R. White) and the integrated approach has also been empirically implemented (see David Backus and Purvis). If we wish to propose our model as a more general basis for estimating a system of asset-demand functions and a consumption function, all of the estimation problems discussed by Smith and Purvis will still be present. Particularly important are the explicit endogeneity of X_m (causing dependence between X_m and the stochastic disturbance terms) and the possibility of simultaneity problems not explicitly considered (for example, to assume that income is predetermined may be unrealistic). Severe multicollinearity is invariably present (although Smith and Brainard have shown that this problem is not insurmountable). The singularity of the variance-covariance matrix of contemporaneous disturbance terms and the non-linear nature of our adding-up restrictions will also be important considerations in estimation.

REFERENCES

D. Backus and D. Purvis, "An Integrated Model of Household Flow-of-Funds Allocations," *J. Money, Credit, Banking*, May 1980, *12*, 400–21.

W. C. Brainard and J. Tobin, "Pitfalls in Financial Model Building," *Amer. Econ. Rev. Proc.*, May 1968, *58*, 99–122.

B. M. Friedman, "Financial Flow Variables and the Short-Run Determination of Long-Term Interest Rates," *J. Polit. Econ.*, Aug. 1977, *85*, 661–89.

R. W. Kopcke, "U.S. Household Sector Demand for Liquid Financial Assets, 1959–1970," *J. Monet. Econ.*, Oct. 1977, *3*, 409–41.

D. D. Purvis, "Dynamic Models of Portfolio Behavior: More on Pitfalls in Financial Model Building," *Amer. Econ. Rev.*, June 1978, *68*, 403–9.

M. Saito, "Household Flow-of-Funds Equations: Specification and Estimation," *J. Money, Credit, Banking*, Feb. 1977, *9*, 1–20.

G. Smith, "Pitfalls in Financial Model Building: A Clarification," *Amer. Econ. Rev.*, June 1975, *65*, 510–6.

______, "Dynamic Models of Portfolio Behavior: Comment on Purvis," *Amer. Econ. Rev.*, June 1978, *68*, 410–16.

______ **and W. Brainard,** "The Value of A Priori Information in Estimating a Financial Model," *J. Finance*, Dec. 1976, *31*, 1299–1322.

W. R. White, "Some Econometric Models of Deposit Bank Portfolio Behaviour in the UK, 1963–70," in G. A. Renton, ed., *Modelling the Economy*, London 1975.

The Choice of Discount Rates for Public Projects

By ROBERT MENDELSOHN*

For two decades, economists have debated whether the future cost and benefits of public projects should be discounted by a social rate of time preference or a marginal rate of return on private investment.[1] It is apparent from the literature produced by this debate that consumption in each period should be discounted by the social rate of time preference. Investments are treated as merely streams of future consumption. The correct discount rate thus hinges upon the social present value of a dollar of private investment (V_t), which is defined as the value of the stream of consumption from the investment discounted by the social rate of time preference. Is the social value of a dollar of private investment greater than a dollar of consumption? David Bradford addresses this issue with a sophisticated analysis of the stream of consumption from private investment. His conclusion, in a wide variety of circumstances, is that the social present value of a dollar of private investment is worth little more than a dollar of consumption. Bradford therefore advocates the widespread adoption of a social rate of time preference to evaluate all public projects.

Although Bradford's synthesis of the literature on public discount rates is impeccable, his calculations of the social present value of a dollar of private investment (V_t) are flawed. Bradford confuses society's marginal propensity to save from capital (S_k). When corrected, Bradford's model indicates that the social present value of a dollar of private investment is often worth considerably more than a dollar of today's consumption. It is clearly not a wise rule of thumb to use the social rate of time preference as the public discount rate.

The critical issue in this paper is Bradford's calculation of the social present value of a dollar of private investment (V_t). An investment is construed to last only one period and it yields the market rate of return $(1+r)$. Part of this return is consumed at the end of the period and the remainder is reinvested. Multiperiod investments are captured by this model through a recursive process. The beauty of this formulation is that it captures the reinvestment decisions of investors. Each period the investor decides what fraction of his capital (S_k) he wishes to reinvest. If all of the above parameters are constant throughout time, the present value of a dollar of private investment is

(1)

$$V_t = \frac{(1+r)(1-S_k)}{1+i} + \frac{(1+r)^2(1-S_k)S_k}{(1+i)^2} + \dots \frac{(1+r)^n(1-S_k)S_k^{n-1}}{(1+i)^n}$$

Bradford, unfortunately, misinterprets the fraction of capital reinvested (S_k) as the fraction of income that is reinvested (S_y). Thus, Bradford's suggestion of a savings rate of 10–30 percent, while appropriate for S_y, is much too low for S_k. At these low savings rates, the representative capitalist actually consumes between 65 and 89 percent of his capital stock each year. Although investors may be saving only a small fraction of their incomes, they are clearly saving enough to maintain their capital.

The correct value for the marginal propensity to save from capital can be calculated from the following formula:

$$S_k = \frac{1+r\cdot S_y}{1+r} \tag{2}$$

Thus, if the marginal propensity to save

*University of Washington.

[1] Among the most notable contributions to this debate are Otto Eckstein; Stephen Marglin (1963a,b); Kenneth Arrow; Arrow and Mordecai Kurz.

TABLE 1—THE SOCIAL PRESENT VALUE OF A DOLLAR OF PRIVATE INVESTMENT (V_t)[a]

	$S_y=0$			$S_y=.10$		
r	.05	.10	.15	.05	.10	.15
i						
.02	2.5	5.0	7.5	3.0	9.0	27.0
.05	1.0	2.0	3.0	1.0	2.3	3.9
.08	.6	1.3	1.9	.6	1.3	2.1
	$S_y=.20$			$S_y=.30$		
r	.05	.10	.15	.05	.10	.15
i						
.02	4.0		∞	7.0	∞	∞
.05	1.0	2.7	6.0	1.0	3.5	21.
.08	.6	1.3	2.4	.5	1.4	3.

[a] I thank Robert Halvorsen for noting that the figures above can be reproduced with the formula:

$$V_t=\frac{(1-S_y)r}{i-S_y-r}$$

from income is between .10 and .30 and the market interest rate is .10, the marginal propensity to save from capital is between .918 and .936. Bradford's substitution of the marginal propensity to save from income (S_y) for the marginal propensity to save from capital results in sizeable miscalculations for the value of V_t.

Suppose an investor buys a perpetuity which provides a constant return each period ad infinitum and he consumes this return each period.[2] As long as taxes on income from capital have driven a wedge between the market interest rate and the social rate of time preference, the present value of this dollar of private investment is worth more than a dollar (see Table 1). For example, if the market rate of interest is 5 percent and the social rate of time preference is 2 percent, the present value of the future consumption stream is 2.5. Even without reinvestment, the social present value of a dollar of private investment is clearly greater than a dollar of consumption.

The present value of a dollar of investment is worth even more if the future returns from that investment are reinvested. For example, suppose the person invests one dollar in period 0 which then earns r each period forever. In the second period, the person receives r income. He splits this income into further investment $r \cdot s$ and consumption $r \cdot (1-s)$. Next period he receives $r+r^2s$ income and splits these funds into consumption $(1-s)(r+r^2s)$ and investment $s(r+r^2s)$. The consumer continues in this manner forever. The present value of a dollar of investment is the discounted value of the resulting consumption stream using the social rate of time preference as the discount rate. These calculations are shown in Table 1 for a range of values for the savings rate of income (S_y), the market rate of interest, and the social rate of time preference.

As is clearly evident in Table 1, the social present value of a dollar of private investment varies considerably depending upon the parameters. When the market interest rate equals the social rate of time preference, V_t is equal to one. A dollar of private investment is then equal to a dollar of consumption. In contrast, V_t can be infinite when the market interest rate is considerably greater than the social rate of time preference and the savings rates are high. The consumption stream, in this case, is growing at a faster rate than the social rate of time preference. Diversion of private in-

[2] If the market interest rate was 10 percent, this investor would consume his ten cents of net income in the first period and reinvest the remaining dollar. His reinvestment rate from capital at the end of the period would be about .90, his reinvestment rate from income (S_y) would be zero.

vestments into public projects which yield the social rate of time preference is a serious error in this circumstance. Even with moderate rates of saving from income (.20), discrepancies between the market rate of interest and the social rate of time preference result in values of V_t which are considerably larger than one. For example, if r is equal to .15 and i is equal to .05, one dollar of private investment is worth six dollars of consumption.

The social present value of a dollar of private investment is quite sensitive to 1) the market interest rate, 2) the social rate of time preference, and 3) the rate of savings from income. Consequently, no single discount rate can act as a satisfactory rule of thumb under all circumstances. If the social rate of time preference is used as the public discount rate, the opportunity cost of public investment would be understated, oftentimes by a factor of three or more. Conversely, the market rate of interest tends to overstate the opportunity cost of funds taken from the private sector. Although in practice it may be quite difficult to estimate the optimal discount rate for public projects, it is clearly important to find and utilize the appropriate value.

REFERENCES

Kenneth Arrow, "Discounting and Public Investment Criteria," in A. V. Kneese and S. C. Smith, eds., *Water Resources Research*, Baltimore 1966.

——— **and Mordecai Kurz,** *Public Investment, the Rate of Return, and Optimal Fiscal Policy*, Baltimore 1970.

D. Bradford, "Constraints on Government Investment Opportunities and the Choice of Discount Rates," *Amer. Econ. Rev.*, Dec. 1975, *65*, 887–99.

O. Eckstein, "Investment Criteria for Economic Development and the Theory of Intertemporal Welfare Economics," *Quart. J. Econ.*, Feb. 1957, *71*, 56–85.

S. Marglin, (1963a) "The Social Rate of Discount and the Optimal Rate of Investment," *Quart. J. Econ.*, Feb. 1963, *77*, 95–111.

———, (1963b) "The Opportunity Costs of Public Investment," *Quart. J. Econ.*, May 1963, *77*, 274–89.

Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination

By RICHARD SCHMALENSEE*

Under pure Pigouvian third-degree price discrimination, a monopolist maximizes profits by charging different prices to different markets or classes of customers, with no (second-degree discriminatory) bulk discounts or other non-linear pricing allowed. The standard comparison of such conduct with that of a single price monopoly remains that presented by Joan Robinson (Book V) almost a half century ago. Using an algebraic approach, my note generalizes and extends some of her main results.

Robinson (pp. 190–92) shows geometrically that if a single price monopoly selling in two markets under constant costs is allowed to discriminate between them, total output is unchanged if both markets have linear demand curves.[1] This result is easily extended to the *N*-market case below. If demand curves are not linear, she argues (pp. 192–95) that a comparison of their "adjusted concavities" at the nondiscriminating monopoly price determines whether total output rises or falls. Her formal argument depends critically on the assumption that the discriminating monopoly's prices are nearly equal, however, and M. L. Greenhut and H. Ohta show by (non-pathological) example that her proposed test does not work when those prices differ substantially.[2] The general relation between curvature of demand functions and total output changes due to monopolistic discrimination is analyzed below.

Robinson's discussion of the welfare implications of discrimination (ch. 16) is very brief and informal, and it emphasizes equity as much as efficiency. Perhaps because of this, many subsequent authors seem to equate the efficiency effects of discrimination with its impact on total output.[3] Basil Yamey uses a rather special example to argue that this equation is invalid; he asserts that in general the usual Marshallian welfare measure falls unless output increases. This assertion is confirmed below for the general *N*-market case with arbitrary demand function curvatures. As an immediate corollary, it follows that if all demands are linear, prohibiting a monopoly from practicing third-degree discrimination produces a net welfare gain.[4]

The intuition behind these last results was presented by A. C. Pigou (pp. 284–85, 288–89) and cited by Robinson (p. 206). For any fixed total output of the monopolized product, efficiency requires that all buyers have the same marginal valuation of additional units. (If all buyers are households, they must have the same marginal rate of substitution between the good involved and any numeraire good.) Selling the same product

*Professor of applied economics, Sloan School of Management, Massachusetts Institute of Technology. I am indebted to Vincent Crawford, Jeffrey Harris, and Robert Pindyck for useful comments.

[1] It is important to point out, as Robinson (pp. 188–90) does but many subsequent authors do not, that this result depends critically on the assumption that both markets are served under both regimes. In general, the profit-maximizing nondiscriminatory price may be so high that no purchases are made in markets that would be profitably served under discrimination. If this occurs and demands are linear, allowing discrimination serves to *increase* total output by exactly the amount sold in the previously excluded markets. (See Merton Miller and Raymond Battalio and Robert Ekelund, Jr., for more on such cases.) On the other hand, John Kwoka, Jr., has recently shown by example that if such exclusion can occur, allowing a monopoly to practice second-degree discrimination (in the form of declining-block pricing) can *reduce* total output when demand curves cross.

[2] The related local tests of Edgar Edwards and Thomas Finn share this same defect.

[3] For a recent example, see Robert Bork, pp. 394–398.

[4] This result and that cited in the preceding sentence require that the same markets be served under both regimes (see fn. 1) and that distributional and income effects be neglected, so that the aggregate Marshallian surplus has welfare content.

at different prices to different buyers induces different marginal valuations and produces what Robinson terms "a maldistribution of resources as between different uses" (p. 206). Only an increase in total output above the single price monopoly level can serve to offset this distributional inefficiency. Thus, unless total output increases, monopolistic third-degree price discrimination produces a net efficiency loss.

I. Preliminaries

Consider a monopolist selling in N distinguishable markets (or to N distinguishable customer classes). Let q_i be unit sales in market i, let Q be the sum of the q_i, and let p_i be the price charged in market i. For simplicity, it is generally assumed to be optimal for the monopolist to make positive sales in all N markets, whether or not discrimination is possible. (See fn. 1.) Following the relevant literature (for reasons discussed below), q_i is assumed to depend only on p_i for $i=1,\dots,N$, and unit cost c is assumed constant.

The monopoly's total profit can be written as

$$\text{(1)}\quad \Pi=\sum_{i=1}^{N}(p_i-c)q_i(p_i)\equiv\sum_{i=1}^{N}\pi_i(p_i)$$

where $q_i(p_i)$ is the demand function for market i, and $\pi_i(p_i)$ is net profit generated in that market, for $i=1,\dots,N$. It is assumed that the π_i are smooth, strictly concave functions. (We basically need smooth and declining marginal revenue curves.) If discrimination is impossible, profits are maximized by charging all buyers p^*, the unique solution to

$$\text{(2)}\quad \sum_{i=1}^{N}\pi_i'(p^*)=\sum_{i=1}^{N}\left[(p^*-c)q_i'(p^*)+q_i(p^*)\right]=0$$

On the other hand, if pure third-degree discrimination is possible, the N optimal prices are found as the unique solutions to

$$\text{(3)}\quad \pi_i'(p_i^*)=\left[(p_i^*-c)q_i'(p_i^*)+q_i(p_i^*)\right]=0 \qquad i=1,\dots,N$$

Following Robinson's terminology, let the *strong* markets be those in which p_i^* exceeds p^*, and let S be the set of the corresponding indices. Similarly, let W be the set of indices of the *weak* markets, in which $p^*>p_i^*$, and let I be the set of indices of the *intermediate* markets, if any, for which $p^*=p_i^*$. Because unit cost is constant and individual market demands are determined only by own prices, it is immediate that $i\in S$ $(i\in W)$ if and only if $\pi_i'(p^*)$ is positive (negative). Under such regular and separable conditions, gradient methods work well, and a related method is used in what follows. It may be possible to permit some cross effects and still use the basic approach employed here, but I have so far found no economically meaningful way of doing this.

If income effects are assumed small and distributional effects are neglected, we can employ the standard aggregate Marshallian welfare indicator, consumer's surplus plus producer's (excess) profits:

$$\text{(4)}\quad W=\sum_{i=1}^{N}\left\{\int_{p_i}^{\infty}q_i(v)\,dv+\pi_i(p_i)\right\}$$

Let us now consider the problem of maximizing the strictly concave function $\Pi(p_i,\dots,p_N)$ subject to the linear constraint

$$\text{(5)}\quad \sum_{i=1}^{N}\pi_i'(p^*)(p_i-p^*)\leqslant t$$

where t is some nonnegative constant. If λ is the nonnegative multiplier associated with the constraint, the first-order necessary conditions, which are also sufficient here, are simply

$$\text{(6)}\quad \pi_i'(p_i)=\lambda\pi_i'(p^*)\quad i=1,\dots,N$$

For large t, the p_i^* satisfy (5), the constraint

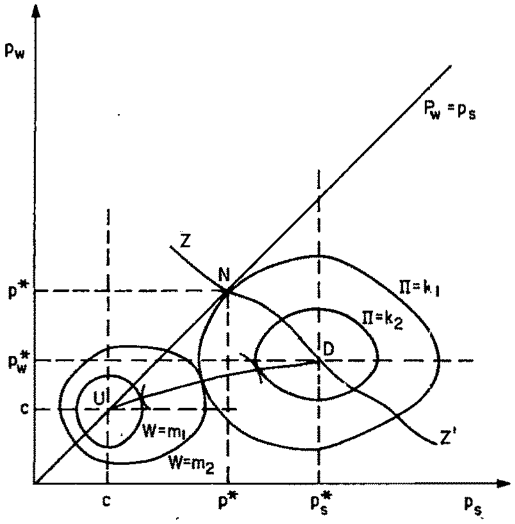


FIGURE 1

does not bind, $\lambda = 0$, and (6) reduces to (3). Suppose that t is small enough so that the constraint is binding and λ is thus positive. Then a reduction in t must increase λ. (If not, π_i' falls (rises) for $i \in S$ ($i \in W$). This in turn raises (lowers) p_i for $i \in S$ ($i \in W$), increasing the left-hand side of (5) and violating the constraint.) When $\lambda = 1$, conditions (6) imply $p_i = p^*$ for all i, and $t = 0$. This problem thus serves to define smooth functions $p_i(t)$ with several useful properties. First, $p_i(0) = p^*$ for all i, and $p_i(t)$ approaches p_i^* as t increases. Second, dp_i/dt has the sign of $\pi_i'(p_i)$ at all points, and it has the sign of $\pi_i'(p^*)$ when both are nonzero. Finally, conditions (2) and (6) imply that the sum of $\pi_i'[p_i(t)]$ is zero for all nonnegative t. Differentiation of that sum yields

$$(7) \quad \sum_{i=1}^{N} [\pi_i''] p_i'(t) = \sum_{i=1}^{N} [2q_i' + (p_i - c)q_i''] p_i'(t) = 0$$

The revelant geometry is illustrated in Figure 1 for a situation with one strong market (p_s) and one weak market (p_w). The point D is the unconstrained optimum, where conditions (3) are satisfied. Because cross-price effects have been assumed away, the iso-profit curves that surround D always have positive slope when $p_w^* < p_w < p^*$ and $p^* < p_s < p_s^*$, as drawn. Point N is the nondiscriminating monopoly optimum, where an iso-profit curve is tangent to the line $p_s - p_w = 0$. The curve ZZ' is the locus of tangencies of such curves with lines of the form $p_s - p_w = T$; its negative slope follows from the strict concavity of the π_i. The functions $p_s(t)$ defined above move prices from N to D along ZZ'. (The rest of Figure 1 is discussed below.)

II. Results

As the parameter t is increased from zero, the functions $p_i(t)$ just defined move the system from single price to discriminating monopoly. We can thus compare output at these two extreme points by using (7) to obtain

$$(8) \quad dQ/dt = \sum_{i=1}^{N} q_i'p_i' = (-1/2)\sum_{i=1}^{N} (p_i - c)q_i''p_i'$$

It follows directly from the second equality that dQ/dt is zero if demand curves are all linear, so that for any N, single price and discriminating monopolies would produce the same total output.

Robinson's (pp. 193–95) "adjusted concavity" test rests on the assumption that the p_i^* are sufficiently close to p^* that, essentially, one need only sign dQ/dt at $t = 0$ in order to determine the effect of monopoly discrimination on total output. In the more natural case in which the p_i^* differ noticeably, so that discrimination suggests itself with some force, it should be clear that this sort of first-order local test can fail, essentially because dQ/dt can change sign. The adjusted concavity test would thus seem to have little real value.

Before presenting that test, however, Robinson (p. 193) makes some general remarks about the global consequences of demand function curvature when $N = 2$ that can readily be verified and extended to the case of $N \geqslant 2$ by examination of (8). If market i is strong, the corresponding term in the second summation in (8) has the sign of $(-q_i'')$; it is thus positive for strictly convex curves and negative for strictly concave ones. Thus if all weak markets have linear demands and all strong market demand curves are strictly convex (concave), a move from single price to discriminating monopoly always raises (lowers) total output, no matter how much the p_i^* differ from p^* and each other. Similarly, strict concavity (convexity) of demand functions in weak markets is associated with output increases (decreases). (Recall that p_i' is always negative for $i \in W$.) If all demand functions are strictly concave or convex and if the p_i^* are not nearly equal, there is apparently no simple, general way to tell if monopolistic discrimination will raise or lower total output.[5]

All the formal analysis so far rests on the assumption that $q_i(p^*)$ and $q_i(p_i^*)$ are positive for all i. This assumption is clearly rather strong, however: some weak markets may not be served at all by a single price monopoly even though a discriminating monopolist could profitably make sales to them. All the results above must therefore be qualified by noting the tendency of a discriminating monopoly to serve markets that would be excluded by a single price seller. The sales made in such markets under discrimination must be added to the output increases computed above in order to assess the full effects of discrimination on total output, a point that Robinson stresses.[6] If one thinks that demand functions are as likely to be concave as convex, recognition of this effect would lead one to conclude that total output is more likely to be increased than decreased by allowing a monopoly to practice third-degree discrimination.

Next, let us go beyond Robinson's analysis and consider the effects of allowing discrimination on the Marshallian welfare measure given by (4).[7] Differentiation of that equation and use of (8) yield

$$(9) \qquad dW/dt = \sum_{i=1}^{N} [p_i - c] q_i' p_i'$$

$$= [p^* - c][dQ/dt] + \sum_{i=1}^{N} [p_i - p^*] q_i' p_i'$$

At $t=0$, $p_i(t) = p^*$ for all i, and the second summation is zero. It is easy to show that it is negative for all $t>0$. If market i is intermediate, the ith term in that summation is zero for all t. But if discrimination causes anything to change, some markets must not be intermediate. If market i is strong (weak), both $[p_i(t) - p^*]$ and $p_i'(t)$ are positive (negative), and the ith term is negative as long as demand slopes down.

Integrating (9) over all nonnegative t implies directly that change in W due to discrimination is always *strictly* less than $(p^* - c)$ times the change in Q. In the linear case, with constant Q, we thus have a drop in Marshallian welfare. In general, unless output increases, movement from single price to discriminating monopoly causes a fall in W, thus a net efficiency loss.

The first term on the right-hand side in (9) resembles the usual expression for the welfare gain from output change in a distorted market: (demand price − marginal cost) × output change. The second term reflects the efficiency cost of distributing total output inefficiently among buyers by driving marginal valuations apart. Equation (9) indicates that the net welfare effect of allowing discrimination is the sum of an output effect of indeterminate sign and a negative distribution effect.

These two effects can be simply illustrated in the two-market case by Figure 2. When discrimination is allowed, price in the strong market rises from p^* to p_s^*, while in the weak market it drops to p_w^*. *In the* strong market, quantity falls by $(q_s^0 - q_s^1)$,

[5] The approach used here does not seem to yield anything of interest when all demand functions have constant elasticities, for instance.

[6] See fn. 1, and the references there cited.

[7] As far as I know, only Yamey has formally considered this measure in the present context, and his treatment is confirmed to an illustrative example that does not explicitly involve third-degree discrimination.

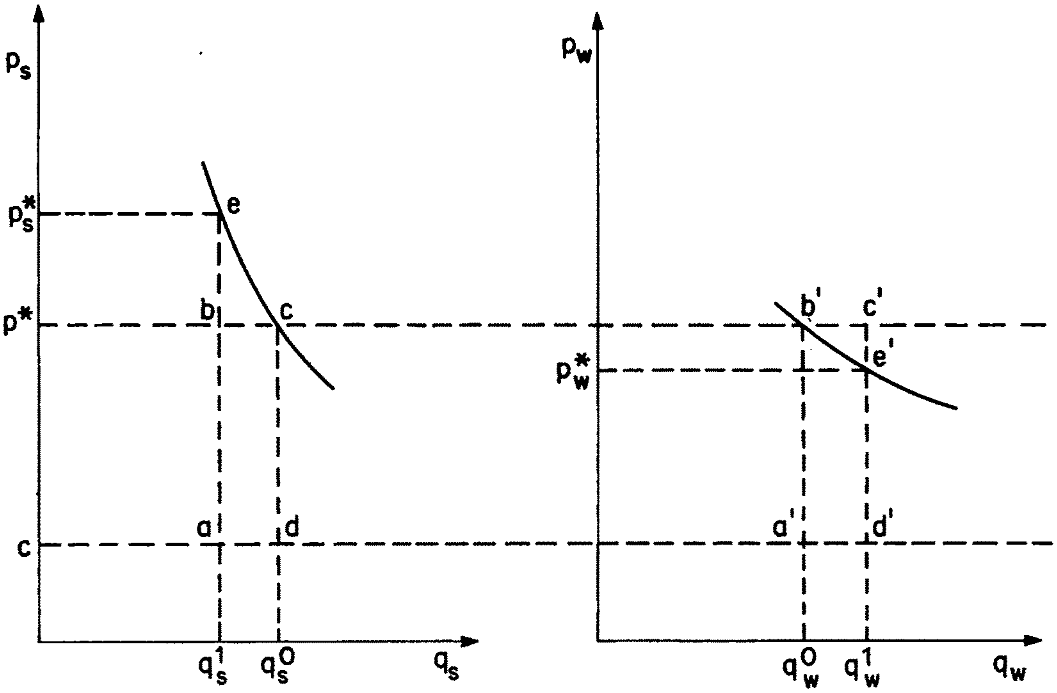


FIGURE 2

while in the weak one it rises by $(q_w^1 - q_w^0)$. The net welfare gain in the weak market is the area $a'b'e'd' = a'b'c'd' - b'c'e'$, while the loss in the strong market is $abcd + bce$. The net gain is thus

$$\Delta W = [a'b'c'd' - abcd] - [b'c'e' + bce]$$

$$= (p^* - c)(Q^1 - Q^0) - (b'c'e' + bce)$$

The net change can thus be positive only if total output expands, only if the increase in sales to the weak market exceeds the drop in sales to the strong market. Clearly $\Delta Q > 0$ is *not* sufficient for $\Delta W > 0$.

If one thinks that demand curves are about as likely to be concave as convex; and if one feels that the Marshallian measure should be taken as seriously as it is taken in most applied welfare analysis, the foregoing discussion might lead one to the conclusion that monopolistic third-degree price discrimination should be outlawed.[8] As before, this must be qualified to some extent by the possibility that such discrimination makes it profitable to sell to markets that would not be served at all under single price monopoly. If discrimination makes possible a large volume of such new sales, it can lead to an increase in welfare even if total sales to previously served markets fail to expand.

Finally, it is worth noting explicitly that nothing here conflicts with the "Ramsey pricing" result that a W-maximizing monopolist subject to a lower bound on Π should practice a milder form of third-degree discrimination.[9] That result is concerned with efficiently trading off welfare against profit, a tradeoff not present in the context of unregulated, profit-maximizing monopoly. In Figure 1, the point U is the unconstrained W-maximizing point. The iso-W loci

[8] It should be clear that such a conclusion would not constitute an endorsement of the Robinson-Patman Act, which cannot fairly be described as simply prohibiting the form of discrimination analyzed here. Also, it is worth reemphasizing the dependence of this analysis on the strong assumption of zero cross-price effects.

[9] See, for instance, William Baumol and David Bradford.

that surround it are easily proven to have negative slope when both prices are above c, as shown. It is clear that any solution to maximizing W subject to a lower bound constraint on Π must lie on the locus of tangencies UD. If the constraint is binding, pricing will involve some degree of discrimination. The point N has no special properties or attraction in this context; nondiscriminatory points generally yield (W,Π) pairs that are dominated by Ramsey points on UD.

REFERENCES

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

R. C. Battalio and R. B. Ekelund, Jr., "Output Change under Third Degree Price Discrimination," *Southern Econ. J.*, Oct. 1972, *39*, 285–90.

Robert J. Bork, *The Antitrust Paradox*, New York 1978.

E. O. Edwards, "The Analysis of Output under Discrimination," *Econometrica*, Apr. 1950, *18*, 163–72.

T. J. Finn, "The Quantity of Output in Simple Monopoly and Discriminating Monopoly," *Southern Econ. J.*, Oct. 1974, *41*, 239–43.

M. L. Greenhut and H. Ohta, "Joan Robinson's Criterion for Deciding Whether Market Discrimination Reduces Output," *Econ. J.*, Mar. 1976, *86*, 96–7.

J. E. Kwoka, Jr., "Output under Second-Degree Price Discrimination," working paper no. 21, Bureau Econ., Federal Trade Commission, Oct. 1979.

M. H. Miller, "Declining Average Cost and the Theory of Railway Rates," *Southern Econ. J.*, Apr. 1955, *21*, 390–404.

A. C. Pigou, *The Economics of Welfare*, 4th ed., London 1932.

Joan Robinson, *The Economics of Imperfect Competition*, London 1933.

Basil Yamey, "Monopolistic Price Discrimination and Economic Welfare," *J. Law Econ.*, Oct. 1974, *17*, 377–80.

An Explanation for the Correlation of Stocks of Nonhuman Capital with Investment in Human Capital

By JOHN W. GRAHAM*

The purpose of this note is to resolve an apparent conflict between the predictions of traditional human capital theory and the available evidence over the correlation of wealth with investment in human capital. In an influential empirical study published in 1972, Samuel Bowles observed that "family income and wealth are obvious candidates for inclusion in the equation predicting years of schooling" (p. S223) and concluded that "measures of family background explain 52 percent of the variance of the years of schooling obtained" (pp. S233–35). But unfortunately this finding does not square well with the predictions of standard human capital models such as those developed by Gary Becker and Yoram Ben-Porath. As David Levhari and Yoram Weiss have noted: "A somewhat surprising aspect of the traditional theory of human capital is that under perfect capital markets, the amount invested and, in particular, the level of schooling is independent of initial wealth" (p. 957).

It has been argued that the source of this conflict between theory and evidence can be traced to the neoclassical assumptions of perfect capital markets and perfect foresight upon which the theories are based. While it may be true that these assumptions are unrealistic, it is the intent of this paper to demonstrate that they are not necessarily the cause of the conflict. Rather, the problem is an inadequate grounding of human capital theory within the framework of utility maximization. A model with all of the usual neoclassical assumptions that is properly developed within a life cycle context clearly indicates that initial stocks of financial and physical wealth (nonhuman capital) affect the optimal investment in human capital.

It is easy to see why it appears that nonhuman wealth does not matter in the standard model. In that model (reconstructed in Section I), the individual accumulates human capital in each period up to the point where the additional costs of acquiring it just cover the additional benefits. Since neither the costs (foregone earnings, interest, etc.) nor the benefits (expected future earnings) depend upon the individual's holdings of nonhuman capital, the investment decision is independent of initial wealth. But since popular opinion and some empirical work hold that differences in nonhuman wealth do tend to perpetuate differences in human wealth, quite rightly this model has been questioned. And the usual scapegoat turns out to be the unrealistic assumptions upon which that theory is based—perfect capital markets in which an individual can borrow unlimited funds to finance current consumption and investment, or perfect foresight with which an individual can anticipate exactly all future earnings. Becker has suggested that introducing capital market imperfections may be the way to resolve the conflict between theory and evidence, while Levhari and Weiss have shown that assets and human capital accumulation are correlated when future returns are uncertain. In Section II, I present a simple model which retains both of these assumptions but demonstrates that stocks of nonhuman capital can affect human capital accumulation. However, unlike the traditional model which assumes a wealth-maximizing decision maker, my model is

*Assistant professor of economics, University of Illinois, Urbana-Champaign. This paper draws upon my doctoral thesis submitted to Northwestern University. For their valuable suggestions on that work, I would like to thank Robert Eisner, Marc Nerlove, and Barry Friedman.

developed for a utility maximizer in a life cycle context.

Within the past few years several papers have appeared which formally join human capital theory with lifetime labor supply within a utility-maximization framework.[1] But in general these papers have not explored the connection between nonhuman wealth and the human capital investment profile. A notable exception is the important theoretical work by Heckman (1976). However, he reaches the surprising conclusion that "increments in initial financial assets do not affect the acquisition of human capital...[and] inequality in bequests does not cause inequality in human capital stocks and wage rates" (1976, p. S23). In Section III of this paper I show that Heckman's result follows directly from a special assumption about the effect of human capital on the efficiency of nonmarket time, an assumption he has dubbed "Michael-neutrality." I demonstrate that in a nonneutral framework Heckman's model would show a correlation between assets and human capital accumulation.

I. The Wealth-Maximization Model

The standard model of human capital accumulation assumes that an individual invests in himself to maximize his discounted lifetime income, or equivalently, human wealth. Following Levhari and Weiss, the simplest possible life cycle human capital model consists of just two periods. The individual begins his two-period life cycle with initial stocks of financial assets A_0, and human capital H_0. For simplicity, assume he ends the two periods as neither debtor nor creditor to the next generation. And of course, human capital is by definition zero at the end of the life cycle. Financial assets yield a rate of return r, and human capital, when employed, yields a rate of return w. There is a perfect financial capital market that allows an individual to select his optimal portfolio of financial and human assets.

Human capital decays exogenously at a geometric rate σ. Additional human capital can be acquired only through a time-intensive process of self-investment. For simplicity, the only input into the production of human capital is an individual's own time. Let K_1 denote the fraction of the first period devoted to the production of human capital $(0 \leqslant K_1 \leqslant 1)$. The value of human capital at the end of the first period, then, is

$$H_1 = H_0(1-\sigma) + F(K_1) \tag{1}$$

The production function $F(\cdot)$ is assumed to exhibit a positive and diminishing marginal product (i.e., $F' > 0$, $F'' < 0$). Time devoted to producing human capital in the second period, K_2, is zero since no returns would accrue before the terminal date.

The choice problem of an individual is to select K_1, time devoted to acquiring human capital, to maximize total wealth, or equivalently, the present discounted value of lifetime income:

$$A_0 + wH_0(1-K_1) + \frac{wH_1}{1+r} \tag{2}$$

where H_1 is defined in equation (1). Because some time is used to acquire human capital in the first period, measured earnings during that period equal $wH_0(1-K_1)$.

Interior maximization of (2) with respect to K_1 occurs when

$$\frac{dF}{dK_1} = (1+r)H_0 \tag{3}$$

or in other words, when time is devoted to acquiring human capital until its marginal return just equals the market interest rate.

From (3) it is easy to derive the effect of a change in the stock of financial assets on the accumulation of human capital: $\partial K_1/\partial A_0 = 0$. This is the standard result that differences in nonhuman wealth do not lead to differences in the accumulation of human capital. If this conclusion appears unsatisfactory, one might amend the problem such that r depends upon A_0 or such that dF/dK_1 is uncertain.

[1] These include Gilbert Ghez and Becker; Alan Blinder and Weiss; James Heckman (1975, 1976) and Harl Ryder et al.

II. The Utility-Maximization Model

Although consumer behavior is usually based upon the assumption of utility maximization, human capital models have traditionally assumed that wealth is the objective to be maximized. Wealth maximization is assumed to be consistent with the overall objective of utility maximization according to the "separation" theorem which holds that consumption and portfolio-selection decisions can be divided into two sequential decisions: first, assets are chosen to maximize lifetime income, and then they are allocated across periods to finance the consumption plan that maximizes lifetime utility. However, Ryder et al., among others, have shown that the separation theorem is invalid when one of the assets in the portfolio is human capital. Consumption plans cannot be separated from investment in human capital when both use the scarce resource of time.

To demonstrate this result, once again assume an individual is endowed with initial stocks of financial and human capital, lives for two periods, and just exhausts his assets by terminal time. The notation is identical to that of the earlier model. What is different here is that the individual not only chooses his investment in human capital, but also simultaneously chooses his lifetime consumption plans. Write the lifetime utility function as[2]

$$(4) \qquad U=u(x_1,l_1)+v(x_2,l_2)$$

where x_1 and x_2 represent consumption of market goods in periods one and two, and l_1 and l_2 represent leisure time in the two periods. The individual buys leisure by reducing work time, so actual income receipts equal $wH_0(1-l_1-K_1)$ in the first period, and $wH_1(1-l_2)$ in the second.

The decision problem of the individual is to choose consumption of goods and time and investment in human capital to maximize lifetime utility (4), subject to the production function for human capital (1) and the lifetime budget constraint

$$(5) \quad A_0+wH_0(1-K_1-l_1)$$
$$+wH_1(1-l_2)/(1+r)-x_1-x_2/(1+r)=0$$

The first-order condition for optimal investment in human capital is

$$(6) \qquad (1-l_2)\frac{dF}{dK_1}=(1+r)H_0$$

Equation (6) says that investment time should be chosen to equate at the margin the cost of foregone current income with the future returns to that investment, where future returns depend upon future labor supply. Unlike equation (3), investment time can no longer be chosen independently of consumption because of the presence of leisure time in the marginal condition. Separation of consumption from investment is now broken.

In this version of the model we can investigate the effect of changing the initial stock of financial wealth on investment in human capital. It is shown in the Appendix that $\partial K_1/\partial A_0 \neq 0$. In general the sign of the effect is uncertain. However if we assume complete separability between market goods and leisure time, then it can be determined that $\partial K_1/\partial A_0<0$; that is, the greater the stock of financial wealth, the less investment in human capital. Intuitively, since leisure is a normal good, an increase in wealth increases consumption of all goods—including second-period leisure. But according to (6), an increase in leisure, which reduces work time, means any given first-period investment in human capital is less profitable, since less time is spent collecting the returns.

Without the strong separability assumption, the sign of $\partial K_1/\partial A_0$ is uncertain. If the individual can substitute goods for leisure, then additional financial wealth may permit him to consume more goods-intensive commodities, which in turn frees more time for work, increases the returns to human capital investment, and may stimulate additional

[2] The results do not appear to be sensitive to the intertemporal separability assumed in equation (4). However, this assumption does simplify the comparative statics considerably.

investment. The important point, however, is that while the sign of the effect is ambiguous, the effect is not necessarily zero. Financial assets do affect investment in human capital.

III. Utility Maximization with Different Learning and Leisure Technologies

In this section I redo the utility-maximization problem when human capital provides nonmarket benefits. Robert Michael investigated the effect of human capital on the efficiency of leisure time. Ben-Porath suggested that human capital might also affect the efficiency of time devoted to additional human capital accumulation (learning time). Heckman (1976) was first to present a model of optimal human capital accumulation in which the efficiency of leisure, learning, and work time were all influenced by the stock of human capital. In this section I rework Heckman's formulation in considerably simpler mathematics and demonstrate a surprising relationship between the leisure-technology parameter and the influence of nonhuman wealth upon the accumulation of human capital. This result explains why Heckman found no theoretical relationship between nonhuman wealth and human capital accumulation even though his problem was developed in a utility-maximization framework.

Ben-Porath has suggested that the accumulation of human capital might not just increase the efficiency of work time, but also the efficiency of time devoted to acquiring that human capital. However, a priori it is unclear how the efficiency of learning time might improve relative to that of work time. We could rewrite the production function from equation (1) as

$$(7)\quad H_1=H_0(1-\sigma)+F\left(K_1H_0^b\right)\qquad b\geqslant 0$$

If $b=0$, (7) reverts to the original formulation in which learning time is unaffected by the stock of human capital. For $b>0$, the efficiency of learning depends upon the current stock of human capital. Ben-Porath investigated three cases. For $0<b<1$, the efficiency of learning time increases with the stock of human capital, but not as quickly as does work time. For $b>1$, the efficiency of learning time increases faster than the efficiency of market time. For $b=1$, human capital equally augments the efficiency of work and learning time. Heckman has dubbed this case "Ben-Porath-neutrality."

Michael has suggested that human capital might also improve the efficiency of leisure time. Analogous to the specification of the learning technology, we can rewrite the utility function in equation (4) as

$$(8)\quad U=u(x_1,l_1H_0^a)+v(x_2,l_2H_1^a)\quad a\geqslant 0$$

For $a>0$, human capital augments the efficiency of leisure time. Heckman has dubbed the case $a=1$ "Michael-neutrality," that is, where human capital equally augments the efficiency of leisure and work time. In other words, when Michael-neutrality prevails, human capital investments return equal reward in both work and leisure time activities.

It is straightforward to resolve the individual's optimization problem when human capital yields nonmarket benefits. The individual maximizes (8) subject to (7) and the budget constraint (5). The complete set of first-order conditions is presented in the Appendix. The marginal condition on investment time can be simplified to

$$(9)\qquad (1-l_2+al_2)\frac{dF}{dK_1}=(1+r)H_0$$

Like equation (6), equation (9) represents the decision rule for choosing optimal learning time. But unlike that equation, it suggests that the addition to the stock of human capital dF/dK_1 should be weighted not only by second period work $1-l_2$, but also by the extent of leisure-time benefits that accrue, al_2. *Ceteris paribus*, the presence of nonmarket benefits increases the incentive to acquire human capital.

Now examine the special case of Michael-neutrality. For $a=1$, equation (9) reduces to

$$dF/dK_1=(1+r)H_0$$

which is the familiar marginal condition generated by the wealth-maximization model. In other words, when the efficiency of work and leisure are equally affected by the stock of human capital, optimal investment time can be chosen independently of consumption of goods and time. Since human capital provides identical benefits no matter how future time is used, the optimal amount of time devoted to human capital accumulation is independent of the allocation of time between market and nonmarket activities.

How is human capital accumulation influenced by initial stocks of nonhuman capital when nonmarket benefits are present? As the Appendix demonstrates, in general nonhuman assets continue to assert a nonzero influence upon human capital accumulation. However, if Michael-neutrality ($a=1$) is assumed, then the correlation between nonhuman wealth and time devoted to human capital accumulation is zero. This result should not be too surprising, because when $a=1$, the decision rule for choosing K_1 reduces to the marginal condition of the wealth-maximization model in which nonhuman wealth is known to have no influence. Since human capital accumulation does not depend upon the allocation of future time between work and leisure, it does not matter that consumption of leisure time is affected by the stock of nonhuman wealth.

IV. Summary and Conclusions

Are differences in initial stocks of nonhuman wealth associated with differences in human capital accumulation, and ultimately differences in human wealth? Casual observation seems to hold that they are, and studies such as Bowles that equate investment in human capital with years of schooling generally find support for a positive correlation. Of course, to the extent that human capital accumulation consists of more than schooling, there still exists little empirical evidence as to the correlation.[3]

Past theoretical work has shown that nonhuman wealth is correlated with human capital accumulation if capital markets are imperfect or if returns to investment are uncertain. What this paper demonstrates is that neither of these conditions is necessary to show that nonhuman wealth matters. Even if capital markets are perfect and individuals possess perfect foresight, the stock of nonhuman wealth makes a difference. Very simply this is because returns from human capital investment depend upon future work intensity, or equivalently, future consumption of leisure, which, in turn, is affected by the stock of nonhuman wealth. Wealth is independent of accumulation only when the returns themselves are independent of the future consumption of leisure. This occurs when either the "separation theorem" holds or when human capital equally augments the efficiency of work and leisure time. It has been demonstrated that the separation theorem is invalid when human capital is one of the assets in the portfolio, but we cannot so quickly dismiss the possibility of neutrality. The influence of human capital on the efficiency of nonmarket time deserves much more investigation.

Appendix A: The Utility-Maximization Model

$$\max_{x_1, x_2, l_1, l_2, K_1} u(x_1, l_1)+v(x_2, l_2)$$

subject to

$$(1+r)A_0+(1+r)wH_0(1-l_1-K_1)$$
$$+wH_1(1-l_2)-(1+r)x_1-x_2=0$$

and $$H_1=H_0(1-\sigma)+F(K_1)$$

The first-order conditions

$$\text{(A1)}\quad \partial u/\partial x_1-(1+r)\lambda=0$$

$$\text{(A2)}\quad \partial v/\partial x_2-\lambda=0$$

$$\text{(A3)}\quad \partial u/\partial l_1-wH_0(1+r)\lambda=0$$

$$\text{(A4)}\quad \partial v/\partial l_2-w[H_0(1-\sigma)+F(K_1)]\lambda=0$$

[3] My forthcoming paper provides some additional evidence that nonhuman capital affects human capital accumulation by examining a cross section of heads of households whose human capital is estimated as the discounted value of expected earnings.

$$\text{(A5)}\quad \lambda((1-l_2)dF/dK_1-(1+r)H_0)=0$$

$$\text{(A6)}\quad (1+r)A_0+(1+r)wH_0(1-l_1-K_1)$$
$$+wH_1(1-l_2)-(1+r)x_1-x_2=0$$

where λ is the marginal utility of wealth. Totally differentiate the system (A1)–(A6) and set $dr=dw=d\sigma=dH_0=0$. The second-order condition for utility maximization requires that the determinant of the 6×6 matrix shown below be negative. Call this value Z.

We want to sign $\partial K_1/\partial A_0$. Solve for dK_1:

$$dK_1=\frac{(1+r)dA_0}{Z}\frac{dF}{dK_1}$$
$$\times\left\{u_{x_1x_1}u_{l_1l_1}-u^2_{x_1l_1}\right\}\left\{v_{x_2l_2}-v_{x_2x_2}wH_1\right\}$$

Therefore,

$$\partial K_1/\partial A_0=\frac{(1+r)\dfrac{dF}{dK_1}}{Z}$$
$$\times\left\{u_{x_1x_1}u_{l_1l_1}-u^2_{x_1l_1}\right\}\left\{v_{x_2l_2}-v_{x_2x_2}wH_1\right\}$$

In general, this expression cannot be signed. However if we impose intratemporal separability, so that $u_{x_1l_1}=v_{x_2l_2}=0$, and diminishing marginal utility, so that $u_{x_1x_1}<0$, $u_{l_1l_1}<0$ and $v_{x_2x_2}<0$, then

$$\partial K_1/\partial A_0$$
$$=\frac{-(1+r)\dfrac{dF}{dK_1}}{Z}u_{x_1x_1}u_{l_1l_1}v_{x_2x_2}wH_1<0$$

Appendix B: Utility Maximization with Learning and Leisure Technologies

$$\max_{x_1,x_2,l_1,l_2,K_1}\quad u(x_1,l_1H_0^a)+v(x_2,l_2H_1^a)$$

subject to

$$(1+r)A_0+(1+r)wH_0(1-l_1-K_1)$$
$$+wH_1(1-l_2)-(1+r)x_1-x_2=0$$

and $\quad H_1=H_0(1-\sigma)+F(K_1H_0^b)$

The first-order conditions

$$\text{(A7)}\quad \partial u/\partial x_1-(1+r)\lambda=0$$

$$\text{(A8)}\quad \partial v/\partial x_2-\lambda=0$$

$$\text{(A9)}\quad H_0^a\partial u/\partial l_1H_0^a-(1+r)wH_0\lambda=0$$

$$\text{(A10)}\quad H_1^a\partial v/\partial l_2H_1^a-wH_1\lambda=0$$

$$\text{(A11)}\quad \lambda[(1-l_2)w\partial H_1/\partial K_1-(1+r)wH_0]$$
$$+\frac{\partial v}{\partial l_2H_1^a}l_2aH_1^{a-1}\frac{\partial H_1}{\partial K_1}=0$$

$$\text{(A12)}\quad (1+r)A_0+(1+r)wH_0(1-l_1-K_1)$$
$$+wH_1(1-l_2)-(1+r)x_1-x_2=0$$

The last term in (A11) can be simplified using (A10) and $\partial H_1/\partial K_1=dF/dK_1$, so that

$$\begin{bmatrix} u_{x_1x_1} & 0 & u_{x_1l_1} & 0 & 0 & -(1+r) \\ 0 & v_{x_2x_2} & 0 & v_{x_2l_2} & 0 & -1 \\ u_{x_1l_1} & 0 & u_{l_1l_1} & 0 & 0 & -wH_0(1+r) \\ 0 & v_{x_2l_2} & 0 & v_{l_2l_2} & -\lambda w\dfrac{dF}{dK_1} & -wH_1 \\ 0 & 0 & 0 & \dfrac{dF}{dK_1} & (1-l_2)\dfrac{d^2F}{dK_1^2} & 0 \\ -(1+r) & -1 & -wH_0(1+r) & -wH_1 & 0 & 0 \end{bmatrix}\begin{bmatrix} dx_1 \\ dx_2 \\ dl_1 \\ dl_2 \\ dK_1 \\ d\lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -(1+r)dA_0 \end{bmatrix}$$

(A11) reduces to

$$\lambda(1-l_2)w\frac{dF}{dK_1}-\lambda(1+r)wH_0$$

$$+w\lambda al_2\frac{dF}{dK_1}=0$$

or $(1-l_2+al_2)\dfrac{dF}{dK_1}=(1+r)H_0$

which is equation (9) in the text.

Totally differentiate the system (A7)–(A12) and set $dr=dw=d\sigma=dH_0=da=db=0$, as shown in the matrix below. For economy of notation,

$$u_{l_1'l_1'}=\frac{d^2u}{d(l_1H_0^a)^2}$$

and $$z_1=z_2=\left[\frac{dv}{dl_2H_1^a}aH_1^{a-1}\right.$$

$$\left.+v_{l_2'l_2'}l_2aH_1^{2a-1}-\lambda w\right]\frac{\partial H_1}{\partial K_1}$$

$$z_3=\left[\lambda(1-l_2)w+\frac{dv}{dl_2H_1^a}l_2aH_1^{a-1}\right]\frac{\partial^2H_1}{\partial K_1^2}$$

$$+\left[\frac{dv}{dl_2H_1^a}l_2a(a-1)H_1^{a-2}\right.$$

$$\left.+v_{l_2'l_2'}l_2^2a^2H_1^{2a-2}\right]\left[\frac{\partial H_1}{\partial K_1}\right]^2$$

$$z_4=z_5=(1-l_2)w\partial H_1/\partial K_1-(1+r)wH_0$$

$$z_6=z_7=v_{x_2l_2'}l_2aH_1^{a-1}\partial H_1/\partial K_1$$

The second-order condition for utility maximization requires that the determinant of the 6×6 matrix be negative ($Z<0$).

We want to evaluate the sign of $\partial K_1/\partial A_0$. Solve for dK_1:

$$dK_1=\frac{(1+r)dA_0}{Z}H_0^{2a}\left[u_{x_1x_1}u_{l_1'l_1'}-u_{x_1l_1'}^2\right]$$

$$\times\left\{-z_2\left[H_1^av_{l_2'x_2}-v_{x_2x_2}wH_1\right]\right.$$

$$+z_4\left[v_{x_2x_2}v_{l_2'l_2'}H_1^{2a}-v_{l_2'x_2}^2H_1^{2a}\right]$$

$$\left.-z_7\left[H_1^av_{x_2l_2'}wH_1-H_1^{2a}v_{l_2'l_2'}\right]\right\}$$

Divide both sides by dA_0. The sign of $\partial K_1/\partial A_0$ cannot be signed in general.

Evaluate $\partial K_1/\partial A_0$ at $a=1$. When $a=1$,

by (A10) $$z_2=v_{l_2'l_2'}\frac{\partial H_1}{\partial K_1}l_2H_1$$

$$z_4=-l_2w\frac{\partial H_1}{\partial K_1}$$

Since $w(dH_1/dK_1)-(1+r)=0$ at K_1^* from equation (9) and

$$z_7=v_{x_2l_2'}\frac{\partial H_1}{\partial K_1}l_2$$

Therefore,

$$\partial K_1/\partial A_0=\frac{(1+r)H_0}{Z}\left[u_{x_1x_1}u_{l_1'l_1'}-u_{x_1l_1}^2\right]$$

$$\times\left\{-\left[H_1v_{l_2'x_2}-v_{x_2x_2}wH_1\right]v_{l_2'l_2'}\frac{\partial H_1}{\partial K_1}l_2H_1\right.$$

$$\begin{bmatrix} u_{x_1x_1} & 0 & H_0^au_{x_1l_1'} & 0 & 0 & -(1+r) \\ 0 & v_{x_2x_2} & 0 & H_1^av_{x_2l_2'} & z_6 & -1 \\ H_0^au_{x_1l_1'} & 0 & H_0^{2a}u_{l_1'l_1'} & 0 & 0 & -(1+r)wH_0 \\ 0 & H_1^av_{x_2l_2'} & 0 & H_1^{2a}v_{l_2'l_2'} & z_1 & -wH_1 \\ 0 & z_7 & 0 & z_2 & z_3 & z_4 \\ -(1+r) & -1 & -(1+r)wH_0 & -wH_1 & z_5 & 0 \end{bmatrix}\begin{bmatrix} dx_1 \\ dx_2 \\ dl_1 \\ dl_2 \\ dK_1 \\ d\lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -(1+r)dA_0 \end{bmatrix}$$

$$-\left[v_{x_2x_2}v_{l'_2l'_2}H_1^2 - v_{l'_2x_2}^2 H_1^2 \right] l_2 w \frac{\partial H_1}{\partial K_1}$$

$$\left. -\left[H_1^2 v_{x_2l'_2} - H_1^2 v_{l'_2l'_2} \right] v_{x_2l'_2} \frac{\partial H_1}{\partial K_1} l_2 \right\}$$

Combining like terms, the bracketed expression reduces to zero, so $\partial K_1/\partial A_0 = 0$, which proves the result stated in the text.

REFERENCES

Gary Becker, *Human Capital*, 2d ed., New York 1975.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

A. Blinder and Y. Weiss, "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ.*, June 1976, *84*, 449–72.

S. Bowles, "Schooling and Inequality from Generation to Generation," *J. Polit. Econ.*, May/June 1972, Part 2, *80*, S219–51.

Gilbert Ghez and Gary Becker, *The Allocation of Time and Goods Over the Life Cycle*, New York 1975.

J. Graham, "The Influence of Nonhuman Wealth on the Accumulation of Human Capital," in M. Ali Kahn, ed., *Research in Human Capital and Development*, Vol. 2, forthcoming.

J. Heckman, "A Life Cycle Model of Earnings, Learning, and Consumption," *J. Polit. Econ.*, Aug. 1976, Part 2, *84*, S11–44.

______, "Estimates of a Human Capital Production Function Embedded in a Life-Cycle Model of Labor Supply," in Nestor Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1975.

D. Levhari and Y. Weiss, "The Effect of Risk on the Investment in Human Capital," *Amer. Econ. Rev.*, Dec. 1974, *64*, 950–63.

Robert Michael, *The Effect of Education on Efficiency in Consumption*, New York 1972.

H. Ryder et al., "Training and Leisure over the Life Cycle," *Int. Econ. Rev.*, Oct. 1976, *17*, 651–74.

Managed Float: An Evaluation of Alternative Rules in the Presence of Speculative Capital Flows

By CARLOS ALFREDO RODRÍGUEZ*

During the first half of 1978, the Central Bank of Argentina found itself in the uncommon position of managing the exchange rate in the presence of very high inflation rates (6–13 percent a month) and unusually large speculative capital inflows which contributed to much of the increase in the money supply for the period. Given the unusual rate of reserve accumulation, the authorities, intentionally or not, allowed the exchange rate to devalue by substantially less than the rate of domestic inflation. While the policy helped to reduce the comparative advantage of the traded sector, it did not discourage the capital inflows. Thus the authorities were motivated to impose a limited degree of capital controls. High inflation fueled by unusually high capital inflows is not a common phenomenon in these days where high inflation is associated with a weakening currency. It is, however, a distinct possibility as the Argentine experience showed. In this note, an attempt is made to rationalize those events in terms of a simple structural model which can be adapted for the study of different rules of exchange rate management by the central bank.

Consider an economy where domestic and foreign goods are sufficiently differentiated such that they can be treated as two composite bundles of commodities whose relative price, which I will call the real exchange rate, is given by

$$e=\frac{EP^*}{P}$$

where E=nominal exchange rate, P^*=price of foreign goods (in terms of foreign currency), and P=price of domestic goods (in terms of domestic currency). Being the relative price between the goods the country imports and exports, e is expected to be an important variable in the determination of the actual amounts exported and imported, and therefore of the trade balance surplus T. The other major component of the balance of payments is the capital account surplus C, which, it will be shown, can be assumed to depend negatively on the expected *rate of change* in the real exchange rate. That is, a positive expected value for $(1/e)(de/dt)=\hat{e}$ tends to deteriorate the capital account. The reason for the latter result is as follows. Abstaining, for simplicity, from foreign price changes, and assuming that speculators correctly anticipate price changes, the expected change in e can be decomposed into the difference between the percentage change in the nominal exchange rate $\hat{E}$, and the percentage change in the domestic price level $\hat{P}$, i.e., $\hat{e}=\hat{E}-\hat{P}$. Making the reasonable assumption that domestic nominal rates of interest rise in proportion to the domestic inflation rate, $\hat{P}$ is a good proxy for the nominal return (in terms of domestic currency) that speculators can receive; in terms of foreign currency, the rate of return is therefore $\hat{P}-\hat{E}$, since a devaluation of the currency is equivalent to a capital loss to the foreign investor. Since $\hat{P}-\hat{E}$ is the negative of the rate of change in the real exchange rate, it follows that the capital account surplus is negatively associated with $\hat{e}$:

$$C=C_0-\beta\hat{e}$$

where C_0 is that part of capital flows which is not responsive to short-term rates of return (usually associated with direct investment and foreign aid; the sum of C_0 and the trade balance is here denoted as the basic

*Centro de Estudios Macroeconómicos de Argentina. Part of this paper was written while I was visiting scholar at the Asian Department of the International Monetary Fund. The views expressed here do not necessarily represent those of the Fund. I am indebted to Bijan Aghevli and Slobodan Djajic for helpful comments and discussion.

balance, $B=T+C_0$). The parameter β measures the response of capital flows to rates of return, and the larger β, the more perfect the degree of capital mobility.

The overall balance or rate of reserve accumulation dR/dt is

$$\frac{dR}{dt}=T(e)+C_0-\beta\hat{e} \tag{1}$$

The reader must have noticed that so far I have not referred to the monetary side of the economy. I have tried to focus the analysis on real exchange rate adjustment and have tried to construct the simplest model to do so. By assuming that the nominal exchange rate is fully indexed to the price level, we need not determine this last variable within the model in order to obtain the time-path for the *real* exchange rate. In any event, my model is fully consistent with at least the following version of the monetary side. Assume the demand for money equals a constant fraction of real income $\bar{y}$, which is also constant:

$$M^d=P\cdot k\cdot\bar{y} \tag{2}$$

The supply of money comes from the purchases of reserves and creation of domestic credit (D) by the central bank:

$$\frac{dM}{dt}=E\frac{dR}{dt}+\frac{dD}{dt} \tag{3}$$

Differentiating (2) with respect to time and equating to (3), we obtain the expression for the equilibrium rate of change in the nominal money stock:

$$\hat{M}=\left(\frac{e}{ky}\right)\frac{dR}{dt}+\frac{1}{M}\frac{dD}{dt} \tag{4}$$

It also follows from (2) that the rate of change in P equals that of M:

$$\hat{P}=\hat{M} \tag{5}$$

Given the time path of reserves from equation (1), equation (4) determines the rate of change of the nominal money stock, and equation (5) determines the time path of the price level. As noted above, neither the money stock nor the price level are necessary for analysis of the stability of the alternative rules of exchange rate management.

Returning to equation (1), it was shown that the balance of payments improves with the *level* of real exchange rate, but deteriorates with the *rate of change* in the real exchange rate. This fact poses particular problems when the monetary authorities try to control the path of the real exchange rate in order to prevent "disorderly market conditions," while at the same time aiming at the eventual attainment of external balance (i.e., $dR/dt=0$).

Consider a situation in which e is initially high and constant such that the basic balance ($T+C_0$) is in surplus and there are no speculative capital flows. Since reserves are rising, so must the money supply and the price level; however e is assumed constant which means that the nominal exchange rate is devalued at the same rate as domestic inflation. Stopping the ongoing inflation requires eliminating the balance-of-payments surplus and therefore reducing the *level* of the real exchange rate such that the basic balance is brought down to zero (here I am abstracting from growth so that any rate of monetary increase is inflationary). Two different strategies are open in order to bring about the reduction in the real exchange rate:

The Gradualist Solution: Reduce the rate of devaluation *below* the rate of inflation such that the rate of change in the real exchange rate becomes negative. The rate at which e is allowed to fall may in turn be indexed to

A. the overall rate of accumulation of reserves

B. the level of the basic balance.

The Shock Treatment: A one-step sudden (unannounced) revaluation trying to aim at the equilibrium level of the real exchange rate.

There is a basic problem with implementing the type A gradualistic solution. As the policy of gradual fall in the real exchange rate is perceived by the market, capital starts flowing in, and therefore, even before any substantial fall in e is achieved, the surplus in capital account adds to the surplus in the basic balance such that reserves rise by *more* than before. If the rate of slide in e is tied to reserve accumulation, authorities will start making e fall faster, which will in turn induce even larger capital flows, etc. While not absolutely necessary, the outcome of such a process may take the economy onto an unstable path: an ever-falling real exchange rate with rising basic balance deficits and capital account (speculative) surplus.[1]

The type B gradualist policy does not have the above disadvantage. Adjustment of the rate of slide in e to the level of the basic balance yields a stable outcome since reserve changes due to speculative capital flows are ignored for the purposes of determining the rate of adjustment of the exchange rate. As the real exchange rate falls, the basic balance surplus is reduced and so is the rate of fall in e. Eventually e falls enough to yield a zero basic balance at which point the rule calls for an unchanged e and therefore provides no incentives for speculative capital flows.

The shock policy has a disadvantage in that the authorities *do not know* what the equilibrium level of e is. Unless by chance they hit the correct level with the initial revaluation, further adjustment will be necessary and the possibility of disorderly speculation cannot be ruled out.

An analytical evaluation of the stability properties of the gradulist rules A and B is straightforward. Under rule A, the nominal exchange rate is adjusted by the rate of price inflation minus some proportion α of the rate of reserve accumulation:

$$\hat{E}=\hat{P}-\alpha\frac{dR}{dt} \tag{6}$$

Since $\hat{e}=\hat{E}-\hat{P}$, it follows from (6) that the real exchange rate adjusts according to

$$\hat{e}=-\alpha\frac{dR}{dt} \tag{7}$$

Substituting (7) into (1), we obtain the basic differential equation describing the time path of the real exchange rate:

$$\hat{e}=\left(\frac{\alpha}{\alpha\beta-1}\right)\left[T(e)+C_0\right] \tag{8}$$

Stability requires $\delta\hat{e}/\delta e<0$, or

$$\frac{\delta\hat{e}}{\delta e}=\left(\frac{\alpha}{\alpha\beta-1}\right)\frac{dT(e)}{de}<0$$

Since $dT(e)/de$ is positive, stability requires $\alpha/(\alpha\beta-1)<0$ which in turn requires $0<\alpha<1/\beta$. Clearly, the larger the response of capital flows to the rate of return β, the smaller is the stable range for α. Thus, as β grows larger, the smaller should be the rate at which the real exchange rate is allowed to fall in response to reserve accumulation. On the other hand, if capital flows respond slowly to rates of return (small β), α can be large and therefore the authorities can afford to allow e to fall fast whenever reserves are rising. For practical purposes it should be noted that the parameter β is unknown to the authorities and therefore the possibility of choosing an unstable value of α cannot be ruled out by the mere knowledge of the theoretically permissible range ($0<\alpha<1/\beta$). Notice that a negative α will always yield an unstable outcome and thus the intuitively obvious result that the authorities should not allow e to rise when reserves are rising.

Gradualist policy type B implies the following adjustment rule for the nominal rate:

$$\hat{E}=\hat{P}-\alpha\left[T(e)+C_0\right] \tag{9}$$

or

$$\hat{e}=-\alpha\left[T(e)+C_0\right] \tag{10}$$

which is always stable for a positive α since

$$\frac{\delta\hat{e}}{\delta e}=-\alpha\cdot\frac{dT(e)}{de}<0$$

None of the above results are modified if, as can be reasonably expected, the trade balance responds slowly to changes in the

[1] Formally, the dynamic adjustment process just described corresponds to the analytical model below where speculators compute the expected rate of return according to an "adaptive" process.

real exchange rate. Slow adjustment in the trade balance is captured by the following process:

$$(11)\qquad \frac{dT}{dt}=\gamma[T(e)-T]$$

Here $T(e)$ is the "desired" trade balance to which the actual trade balance adjusts with a speed of adjustment of $\gamma>0$. Under policy type A, e and T now adjust according to

$$(12)\qquad \frac{de}{dt}=e\left(\frac{\alpha}{\alpha\beta-1}\right)(T+C_0)$$

$$\frac{dT}{dt}=\gamma[T(e)-T]$$

Around the steady state ($T=T(e)=-C_0$), local stability requires the characteristic roots of the matrix A describing the linear approximation to (12) to have negative real parts. Such a matrix is

$$A=\begin{vmatrix} 0 & e\left(\frac{\alpha}{\alpha\beta-1}\right) \\ \gamma\frac{dT(e)}{de} & -\gamma \end{vmatrix}$$

and its characteristic roots will have negative real parts provided $-\gamma<0$ (guaranteed by assumption) and

$$-\gamma\frac{dT(e)}{de}\cdot e\cdot\left(\frac{\alpha}{\alpha\beta-1}\right)>0$$

which, as before, requires $0<\alpha<1/\beta$.

Under gradualist policy type B the dynamic system is described by

$$(13)\qquad \frac{de}{dt}=-\alpha.e(T+C_0)$$

$$\frac{dT}{dt}=\gamma[T(e)-T]$$

for which the matrix describing the linear approximation around the steady state is

$$B=\begin{vmatrix} 0 & -\alpha e \\ \gamma\frac{dT(e)}{de} & -\gamma \end{vmatrix}$$

This describes a stable system provided $\gamma>0$ (guaranteed by assumption) and $\alpha e\gamma\cdot(dT(e)/de)>0$, which is also satisfied for $\alpha>0$.

The basic model could be criticized on the grounds that speculators are assumed to anticipate perfectly changes in the real exchange rate rather than slowly learning about those changes, possibly according to an adaptive process. The main conclusions are not affected, however, by such modifications. Denote f to be the expected rate of change in the real exchange rate and assume f is revised according to the adaptive process:

$$\frac{df}{dt}=\varepsilon(\hat{e}-f)\qquad \varepsilon>0$$

In the gradulist policy type A, the basic equations describing the motion of the system are now

$$(16)\qquad \frac{dR}{dt}=T(e)+C_0-\beta f$$

$$(17)\qquad \frac{df}{dt}=\varepsilon(\hat{e}-f)$$

$$(18)\qquad \hat{e}=-\alpha\frac{dR}{dt}$$

Substituting (18) into (16) and (17), we have

$$\hat{e}=-\alpha T(e)-\alpha C_0+\alpha\beta f$$

$$\frac{df}{dt}=-\alpha\varepsilon T(e)-\alpha\varepsilon C_0+\varepsilon(\alpha\beta-1)f$$

The two equations above are locally stable provided

$$-\alpha\frac{dT(e)}{de}+\varepsilon(\alpha\beta-1)<0$$

While this condition is clearly less stringent than the previous ones, it is still the case that a large α together with a large β can make the system unstable.

If e is adjusted according to the basic balance, the two differential equations describing the motion of e and f become

$$\hat{e} = -\alpha T(e) - \alpha C_0$$

$$\frac{df}{dt} = -\varepsilon[\alpha T(e) + \alpha C_0 + f]$$

The reader can easily verify that the two equations above are always locally stable independent of the values of α or ε insofar as both are positive.

In conclusion, I have found that a rule allowing the real exchange rate to fall in proportion to the basic balance surplus is stable in spite of the presence of speculative capital flows. Tying the slide of the real exchange rate to the overall rate of reserve accumulation may yield an unstable outcome depending on the relationship between the speeds of response of the monetary authorities and of speculative capital flows; the larger is the speed of response of capital flows, the slower should the real exchange rate be allowed to fall in response to reserve increases. The above results were proved for the case where speculators correctly anticipate changes in the real exchange rate and those expected changes are formed according to an adaptive process; the results also stand up when the trade balance is assumed to react slowly to changes in the real exchange rate.

NOTES

The Secretary of the American Economic Association wishes to announce the results of the mail balloting for officers which took place during the fall of 1980. The following individuals took office January 1, 1981: President-Elect Gardner Ackley; Vice Presidents Otto Eckstein and Alice Rivlin; Executive Committee members Elizabeth E. Bailey and Robert J. Gordon.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada, that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

The Murphy Institute of Political Economy at Tulane University announces the establishment of a Visiting Scholar Program whose purpose is to encourage innovative research in political economy. The position carries no teaching duties and requires residence of one or two semesters at the Tulane campus. The program is limited to individuals who have achieved scholarly distinction, and preference will be given to candidates whose primary research interests are in public sector economics. The stipend is open, depending upon qualifications and length of residency. Interested parties for the academic year 1981-82 should submit their vita and a brief statement of their proposed research to William H. Oakland, Director, Murphy Institute for Political Economy, Tilton Hall, Tulane University, New Orleans, Louisiana 70118.

The 1981 Conference of the International Institute of Public Finance will be held in September 1981 in Tokyo, Japan. The general issue of the meeting is Growth, Inflation, and Public Finance. Suggestions for topics and speakers are invited as are proposals for papers. They may be sent to the President of the International Association, Professor Horst Claus Recktenwald, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Lange Gasse 20, 8500 Nürnberg, West Germany (telephone 0911/5302-200).

The Center for Family Business, with the sponsorship of University Services Institute, Cleveland, Ohio, announces the institution of the Léon A. Danco Award for outstanding research papers in the area of growth and continuity of family-owned companies or closely held corporations. Two awards will be given. One award of $1,000 will be made for a doctoral dissertation. The second award of $750 will be granted for a research paper submitted by candidates for the M.B.A. degree. To be considered for the 1981 awards, the dissertation or research paper must be submitted by the faculty advisor by May 1, 1981, and accompanied by a short letter from the advisor summarizing the main purpose and results of the research. Submit to Donald J. Jonovic, 5862 Mayfield Road, P.O. Box 24268, Cleveland, Ohio 44124.

The *Journal of the Proceedings of Symposia* at the College of Staten Island (*JPS*) would be pleased to review articles on "Technology: Its Effects on Society and the Quality of Life," and "The China Connection" for inclusion in future issues. The deadline is June 15, 1981. Please submit copy to Professor Rosalie Reich, Editor, *JPS*, The College of Staten Island, 715 Ocean Terrace A-324, Staten Island, NY 10301.

The University of Connecticut in conjunction with the U.S. Department of Labor will award $5,000 stipends to individuals who will undertake research on any aspect of workmen's compensation programs. Faculty members or graduate students at the thesis writing stage are eligible to apply. A $2,000 grant will be made to the researcher's university. Proposals received by March 15, 1981 will be evaluated by *April 1981*.

Send proposals to Peter Barth, Department of Economics, U-63, The University of Connecticut, Storrs, CT 06268 (telephone 203+486-3023).

The Faculty Exchange Center was established in 1973 in order to facilitate the movement of professional persons in education. The Center makes it possible for interested professors to exchange positions with colleagues in their field from institutions both on this continent and overseas where the language of instruction is English. In addition to its annual directory devoted to teaching exchanges on the college and university level, the Center will provide two house-exchange supplements, to be published and distributed in spring and fall of 1981. For further information and registration forms, send a stamped, self-addressed envelope to Faculty Exchange Center, 952 Virginia Avenue, Lancaster, Pennsylvania 17603.

The *Cambridge Journal of Economics* is offering a prize of £250 for the best paper on the topic "America in the World Economy," submitted by a graduate student (anyone within five years of obtaining his/her first degree or within five years of registration as a graduate student). Papers must be submitted by June 30, 1981. All persons intending to submit papers should first write for full details and regulations to Sarah Bourne, Managing Editor, *CJE*, Faculty of Economics and Politics, Sidgwick Avenue, Cambridge CB3 9DD, UK.

Call for Papers: The tenth edition of *Socioeconomic Issues of Health* will be published in December 1981. This volume is devoted to public policy issues related to the U.S. health care system such as regulation, national health insurance, and health maintenance organizations. Abstracts should be sent no later than May 1, 1981, and papers no later than July 1, 1981, to Editor, *Socioeconomic Issues of Health*, Center for Health Services Research and Development, American Medical Association, 535 North Dearborn Street, Chicago, IL 60610.

The annual meeting of the New England Slavic Association will be held in Portland, Maine, April 24-25, 1981. For full information, contact Professor Frank Durgin, School of Business and Management, University of Southern Maine, Portland, ME 04101.

The International Union for the Scientific Study of Population will hold its nineteenth General Conference in Manila (Philippines), December 9-16, 1981. The program will include the following sessions: Reassessment of Population Trends, From Rome to Manila; How Demography has Changed in Three Decades; Fertility: Trends, Determinants, and Consequences; Fertility and its Regulation, Nuptiality and Family; Mortality, Migration and Population Distribution; Economic Demography: Data Collection and Methodology; Projections; Specific Issues to the Demography of Particular Groups; and informal sessions on various aspects of the population field. For further information and details, contact Bruno Remiche, Executive Secretary, IUSSP, Rue Forgeur 5, 4000 Liège, Belgium.

The twelfth Atlantic Economic Conference will be held in New York City, October 8-11, 1981. Its theme is "A Taste of Excellence." The submission deadline for papers is March 15, 1981. For full information, contact John M. Virgo, Program Chairman, Atlantic Economic Conference, Box 258, Worden, IL 62097.

The North American Economic Studies Association is organizing several sessions for the December 1981 Allied Social Science Association meeting to be held in Washington, D.C. Detailed abstracts of two-to-three pages should be sent by April 15, 1981 to the Secretary-Treasurer and Program Chairman, Dr. Alan M. Rugman, Centre for International Business Studies, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H8.

Marketing Science, a new journal jointly sponsored by the Institute of Management Sciences and Operations Research Society of America, invites quantitatively oriented marketing manuscripts. The papers should make a significant contribution to understanding of marketing phenomena and can be in the form of theory, model, measurement (or estimation) precedure, or application of a new methodology. For full information, contact Editor Donald G. Morrison, Graduate School of Business, 414 Uris Hall, Columbia University, New York, NY 10027. The first issue is expected to appear in January 1982.

Deaths

Edward D. Kalachek, professor of economics, Washington University (St. Louis), Dec. 9, 1979.

Edward F. Meeker, director, department of economic research, Center for Health Services Research and Development, American Medical Association, Chicago, Aug. 14, 1980.

Richard G. Milk, visiting associate professor of economics, Virginia Commonwealth University, May 21, 1980.

Edward J. Powers, professor of economics, Northern Michigan University, Aug. 8, 1980.

Retirement

Nicolas Spulber, professor of economics, Indiana University, July 1, 1980.

Visiting Foreign Scholars

Pekka Ahtiala, University of Tampere, Finland: visiting professor of economics, Northwestern University, spring 1981.

Monojit Chatterji, University of Essex, England: visiting associate professor, department of economics, University of California-Davis, July 1, 1980.

Hajime Hori: visiting associate professor, department of economics, University of Iowa, Aug. 1980.

Rikard Lang, University of Zagreb, Yugoslavia: department of economics, Florida State University, and Center for Yugoslav-American Studies, Jan.–Apr. 1981.

Promotions

Marcelle V. Arak: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, July 1, 1980.

Robert T. Falconer: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, July 1, 1980.

William J. Gasser: manager, external financing department, Federal Reserve Bank of New York, July 1, 1980.

Richard J. Gelson: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, July 1, 1980.

R. Jeffery Green: professor of economics, Indiana University, July 1, 1980.

Oskar R. Harmon: assistant professor of economics, Bentley College, Sept. 1980.

Duane G. Harris: professor of economics, Iowa State University, July 1, 1980.

Daphne A. Kenyon: assistant professor of economics, Dartmouth College, July 1, 1980.

Michael A. Klein: professor of economics, Indiana University, July 1, 1980.

Roger M. Kubarych: vice president and assistant director of research, Research and Statistics Division, Federal Reserve Bank of New York, July 1, 1980.

Patricia H. Kuwayama: manager, statistics department, Federal Reserve Bank of New York, July 1, 1980.

Randolph C. Martin: professor of economics, University of South Carolina, 1980.

John A. Miranowski: associate professor of economics, Iowa State University, July 1, 1980.

Stephen M. Renas: professor of economics, Wright State University, Sept. 1, 1980.

Thomas M. Stevenson: professor, department of economics, St. Louis University, July 1, 1980.

Kenneth E. Stone: associate professor of economics, Iowa State University, July 1, 1980.

John A. Wenniger: manager, domestic research department, Federal Reserve Bank of New York, July 1, 1980.

Ronald Wilder: professor of economics, University of South Carolina, 1980.

James E. Zinser: professor of economics, Oberlin College, July 1, 1980.

Administrative Appointments

Larry M. Blair: director, Labor and Policy Studies Program, Oak Ridge Associated Universities, Sept. 1980.

Walter A. Fogel: associate director, Institute of Industrial Relations, University of California-Los Angeles, July 1, 1980.

Philip Friedman: director, MBA Program, Boston University, June 1, 1980.

Sydney S. Hicks, Federal Reserve Bank of Dallas: vice president, investments and funding, First National Bank of Dallas, Feb. 1980.

James T. Little: chairman, Urban Studies Program, Washington University (St. Louis), July 1, 1980.

Shlomo Maital: head, Economics Section, Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, May 1980.

Laurence H. Meyer: chairman, economics department, Washington University (St. Louis), July 1, 1980.

Robert Piron: chairman, department of economics, Oberlin College, Feb. 1, 1981.

Barton Wechsler: acting director, department of economics, Wright State University, June 1980.

John E. Weiler: chairman, department of economics and finance, University of Dayton, Aug. 16, 1980.

Patrick J. Welch: chairman, department of economics, St. Louis University, July 1, 1980.

Appointments

Michael Abrahams: assistant professor, department of economics, Iowa State University, Sept. 1, 1980.

John T. Addison: associate professor of economics, University of South Carolina, 1980.

Klaus F. Alt: assistant professor, department of economics, Iowa State University, July 1, 1980.

Stephen E. Baldwin, Bureau of Labor Statistics: senior staff economist, National Commission for Employment Policy, Sept. 1980.

Thomas A. Barthold: assistant professor of economics, Dartmouth College, July 1, 1980.

Todd A. Behr, Lehigh University: instructor of economics, Colby College, Sept. 1980.

Herminio A. Blanco-Mendoza, Instituto Tecnológico de Mexico: assistant professor of economics, Rice University, July 1980.

William C. Bonifield, Wabash College: dean, College of Business Administration and professor of economics, Butler University, Jan. 1981.

Gordon L. Brady: graduate fellow, Yale Law School and lecturer, economics department, Yale University, Sept. 1, 1980.

Meyer Burstein: visiting professor, economics department, University of Miami, Aug. 1980.

Thomas C. Campbell, West Virginia University: visiting professor of economics, Virginia Commonwealth University, Aug. 1980.

Edward S. Cavin, University of Michigan: economist, Mathematica Policy Research, Sept. 1980.

Henry W. Chappell, Jr.: assistant professor of economics, University of South Carolina, 1980.

Gregory B. Christainsen, University of Wisconsin: instructor of economics, Colby College, Sept. 1980.

David Cleeton: assistant professor, department of economics, Oberlin College, July 1, 1980.

Richard D. Coe: assistant professor of economics, University of Notre Dame, fall 1980.

David Colander: associate professor, economics department, University of Miami, Aug. 1980.

Gregory Crespi: visiting assistant professor, department of economics, University of Iowa, Aug. 1980.

Roger A. Dahlgran: assistant professor, department of economics, Iowa State University, Apr. 1, 1980.

Ronald E. Deiter: assistant professor, department of economics, Iowa State University, July 1, 1980.

William R. Dougan: instructor, Dartmouth College, July 1, 1980.

Robert A. Driskill: assistant professor of economics, University of California-Davis, July 1, 1980.

Steven Dym: economist, Business Conditions Division, domestic research department, June 4, 1980.

Barry L. Falk: instructor, department of economics, Iowa State University, Sept. 1, 1980.

Luis F. Fernandez: instructor, department of economics, Oberlin College, July 1, 1980.

Rudy Fichtenbaum: assistant professor of economics, Wright State University, Sept. 1, 1980.

Gary Gappert: professor of urban studies and director, Institute for Future Studies and Research, University of Akron.

J. Fred Giertz, Miami University: professor of economics, Institute of Government and Public Affairs, University of Illinois, Aug. 1980.

Jean B. Grossman, Massachusetts Institute of Technology: economist, Mathematica Policy Research, July 1980.

Dennis Hall-Martindale: lecturer, department of economics, Dartmouth College, 1980-81.

William T. Harris: associate professor of economics, Louisiana Tech University, Sept. 1980.

Bruce Herrick, University of California-Los Angeles: head, department of economics and professor of economics, Washington and Lee University, Sept. 1980.

Bryon Higgins: visiting assistant professor, department of economics, University of Iowa, Jan. 1981.

Hillard G. Huntington, Data Resources, Inc.: senior research associate, Energy Modeling Forum, Stanford University, Sept. 1980.

Tatsuro Ichiishi: associate professor, department of economics, University of Iowa, Aug. 1980.

Sanford Jacoby: assistant professor, Graduate School of Management, University of California-Los Angeles, July 1, 1980.

Robert T. Jerome, Jr.: visiting instructor of economics, University of Notre Dame, 1980-81.

Dennis Johnson: visiting professor, department of economics, University of Iowa, Aug. 1980.

David A. Katz: associate professor of economics, University of Dayton, Aug. 16, 1980.

A. Wahhab Khandker: assistant professor of economics, Wabash College, Aug. 1980.

Sheldon Kimmel, University of Chicago: economist, Antitrust Division, U.S. Department of Justice, Oct. 1980.

Sam L. Lanfranco: visiting assistant professor of economics, Virginia Commonwealth University, Aug. 1980.

Stanley Long: visiting associate professor, department of economics, University of Iowa, Aug. 1980.

Wesley H. Long: commissioner, District of Columbia Public Service (Utilities) Commission, Mar. 31, 1980.

Donald McCloskey: professor of economics and history, University of Iowa, Aug. 1980.

Thomas M. McDevitt: visiting assistant professor, department of economics, Dartmouth College, 1980-81.

Jerome L. McElroy: visiting associate professor of economics, University of Notre Dame, 1980-81.

Lionel W. McKenzie, University of Rochester: research professorship, Harvard University, July 1, 1980–June 30, 1981.

B. Starr McMullen: assistant professor, department of economics, Oregon State University, Sept. 1980.

Forrest Nelson: associate professor, department of economics, University of Iowa, Aug. 1980.

Morley Z. Nkosi: assistant professor of economics, Hofstra University, Sept. 1, 1980.

William Nye, U.S. Department of the Treasury: economist, Antitrust Division, U.S. Department of Justice, June 1980.

Mahboobeh Ordoobadi: visiting assistant professor of economics, Wabash College, Aug. 1980.

Douglas R. Ostrom: lecturer, department of economics, Dartmouth College, 1980-81.

Donald M. Pattillo: assistant professor of finance, University of Dayton, Aug. 16, 1980.

Steve Robinson, William Carey College: professor of economics and coordinator, Division of Business and Economics, Carson-Newman College.

Marius Schwartz, The Brookings Institution: economist, Antitrust Division, U.S. Department of Justice, Oct. 1980.

David Segal: visiting associate professor, department of economics, Dartmouth College, 1980-81.

David Shapiro: visiting professor, economics department, University of Miami, Aug. 1980.

Albert K. Smiley, Princeton University: economist, Antitrust Division, U.S. Department of Justice, Sept. 1980.

Alan H. Smith: visiting associate professor, department of economics, Dartmouth College, 1980-81.

Ross Starr, University of California-Davis: professor of economics, University of California-San Diego, July 1, 1980.

Francis Tapon, University of Guelph, Canada: economist, Antitrust Division, U.S. Department of Justice, July 1980.

Clifford F. Theis, Boston College: instructor, department of economics, Framingham State College, Sept. 1, 1980.

Leonard Wang: assistant professor of economics, Wright State University, Sept. 1, 1980.

Charles Whiteman: instructor, department of economics, University of Iowa, Aug. 1980.

Joseph A. Whitt, Jr.: assistant professor of economics, University of South Carolina, 1980.

Peggy Wier, University of Rochester: economist, Antitrust Division, U.S. Department of Justice, Dec. 1980.

Willard E. Witte, Pennsylvania State University: visiting assistant professor of economics, Indiana University, Aug. 1980.

Leaves for Special Appointment

Robert E. Christiansen, Colby College: visiting Fulbright senior lecturer, University of Malawi.

Henry G. Demmert, University of Santa Clara: economist, Antitrust Division, U.S. Department of Justice, Sept. 1980.

Walter Enders, Iowa State University: visiting professor, McGill University, Sept. 1, 1980–May 31, 1981.

Arnold M. Faden, Iowa State University: visiting scholar, University of California-Berkeley, Sept. 1, 1980–May 31, 1981.

Karl A. Fox, Iowa State University: U.S. Bureau of the Census, July 1, 1980–June 30, 1981.

Roy J. Gardner, Iowa State University: Center for Mathematical Studies in Economics and Management Science, Northwestern University, Sept. 1, 1980–May 31, 1981.

Wallace E. Huffman, Iowa State University: visiting fellow, Economic Growth Center and department of economics, Yale University, Sept. 1, 1980–June 30, 1981.

Philip Jacobs, University of South Carolina: Hospital Cost and Utilization Project, National Center for Health Services Research, 1980-81.

Randall R. Kincaid, Davidson College: Environmental Protection Agency, Washington, 1980-82.

John S. Scott: Bureau of Economics, Federal Trade Commission.

Mary E. Scovill, University of Notre Dame: USAID, U.S. Department of State, Jan. 1980–Dec. 1981.

Abdelaleem H. Sharshar, Virginia Commonwealth University: economist, United Nations Development Program, Jakarta, Indonesia, 1980-81.

Steven M. Sheffrin, University of California-Davis: Brookings economic policy fellow, U.S. Department of the Treasury, Office of the Secretary, 1980-81.

J. Marvin Skadberg, Iowa State University: Land-O-Lakes, Minneapolis, Minnesota, July 1, 1980–June 30, 1981.

George Sweeney, Vanderbilt University: economist, Antitrust Division, U.S. Department of Justice, Jan. 1981.

Leon Wegge, University of California-Davis: visiting research professor, Institute for Econometrics and Operations Research, University of Bonn, Germany, 1980-81.

Kenneth J. White, Rice University: visiting associate professor, Australian National University, spring 1981.

Resignations

Robert L. Avinger, associate professor of economics, Davidson College, June 30, 1980.

Stanley M. Besen, Rice University: senior economist, Rand Corporation, Washington, Sept. 1980.

Donald L. Martin, University of Miami: senior consultant, National Economic Research Associates, Inc., New York City.

Michael Podgursky, University of Notre Dame: University of Massachusetts, fall 1980.

David Segal, Oberlin College, June 30, 1980.

Caroline Swartz, University of Notre Dame: Emory University, fall 1980.

E. Lane Vanderslice, University of Notre Dame: Bread for the World, fall 1980.

Neil R. Wright, Rice University: American National Bank, Chicago, July 1980.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution, and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Editorial Office, University of California, Los Angeles, CA 90024.

AMERICAN ECONOMIC ASSOCIATION

1981 Annual Membership Rates

Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular member with rank of assistant professor or lower or annual incomes of $14,400 or less $30.00
- Regular member with rank of associate professor or annual incomes of $14,400-$24,000 .... $36.00
- Regular member with rank of full professor or annual income above $24,000 .... $42.00
- Junior member (available to registered students for three years only). Student status must be certified by your major professor or school registrar $15.00
- In Countries other than the U.S.A., Add $5.00 to cover postage.
- Family member (second membership without publications; two or more living at same address) ..... $ 6.00

Please begin my issues with:

□ **March** □ **June** (Includes *Papers and Proceedings*) □ **September** □ **December** (Includes 1981 *Survey of Members*)

First Name and Initial ____ Last Name ____ Suffix ____

Address Line 1 or Attention ____

Address Line 2 ____

Address Line 3 ____

City ____ State or Country ____ Zip/Postal Code ____

PLEASE TYPE OR PRINT INFORMATION ABOVE; DO NOT EXCEED SPACES ALLOWED. DUES PAYABLE IN U.S. CURRENCY ONLY, CASHIER'S CHECK OR INTERNATIONAL MONEY ORDER PREFERRED.

Endorsed by (AEA member) ____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____

____

Authorized Signature ____

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212
U.S.A.

NORTH-HOLLAND ANNOUNCES

NORTH-HOLLAND PUBLISHING COMPANY
52 Vanderbilt Ave., New York, N.Y. 10017/ P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Tax Incidence

Contributions to Economic Analysis, Vol. 134

By WOUTER J. KELLER, Netherlands Central Bureau of Statistics

1980. 350 pages.
US $53.75/Dfl. 110.00
ISBN 0-444-86057-6

This book is concerned with the effects of public-sector activities on market oriented economies. The emphasis is on construction of applied general equilibrium models. Using a step-by-step approach, it provides a *comprehensive* treatment of the incidence of taxes.

CONTENTS: 1. Introduction. 2. Taxes and the Public Sector. 3. Dalton's Law. 4. Household Behaviour. 5. Firm Behaviour. 6. The Basic General Equilibrium Model. 7. The Excess Burden in a General Equilibrium Model. 8. Public Consumption. 9. Disaggregating the Household Sector. 10. The Final General Equilibrium Model. 11. Towards an Operational Model. 12. A Multi-Sector Model for The Netherlands.

Labor Unions and the Inequality of Earned Income

Contributions to Economic Analysis, Vol. 129

By JOHN S. PETTENGILL, University of Virginia, U.S.A.

1980. 348 pages.
US $48.75/Dfl. 100.00
ISBN 0-444-85409-6

This book presents the *first mathematical model* of the labor market capable of making explicit quantitative predictions of how the wage structure (the relationship between wages and ability) will adjust to shifts in the demand or supply for labor of different abilities. These predictions take account of firms' decisions as to how much and what quality of labor to employ; and of workers' decisions as to how much labor to supply and how much education to obtain. The model is used to show that for a wide range of parameter values, labor unions substantially increase the inequality of the long-run equilibrium wage structure (for U.S. data, by at least 5% and perhaps as much as 30 or 40%). The model also shows that for the same range of parameter values, a 1% increase in the effective real minimum wage will in the long run, cause employment to shrink between .8 and 4%.

A Theory of the Multiproduct Firm

Studies in Mathematical and Managerial Economics, Vol. 28

By KENNETH LAITINEN, Bell Laboratories, Murray Hill, N. J., U.S.A.

1980. 240 pages.
US $41.50/Dfl. 85.00
ISBN 0-444-85495-9

CONTENTS: 1. The Problem of the Multiproduct Firm. 2. The Technology of Production. 3. Cost Minimization. 4. Profit Maximization. 5. The Revenue Maximizing Firm. 6. Relationships among Cost Minimization, Revenue Maximization, and Profit Maximization. 7. Rational Random Behavior. 8. Econometric Specification and Estimation. 9. Application to Hypothetical Data. 10. Conclusions and Suggestions for Further Research. Appendices.

Recovery and Redistribution under the NIRA

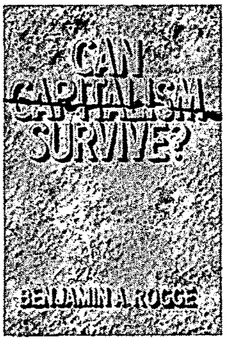
Studies in Monetary Economics, Vol. 6

By MICHAEL M. WEINSTEIN, Haverford College, Pa., U.S.A.

1981. 198 pages.
US $35.50/Dfl. 75.00
ISBN 0-444-86007-X

This book applies modern economic theory and quantitative methods to the analysis of the macroeconomic and redistributive impacts of the National Industrial Recovery Act. This two-year experiment, passed in 1933, was one of the most innovative acts of the New Deal to regulate the U.S. economy. By promulgating industrial codes, economy-wide wage and price controls were imposed, in order to promote recovery, redistribute income, eliminate spiralling deflation, promote social reforms and spread employment. The volume provides insights into the impact of these sweeping wage and price controls in a depressed economy and thus contributes to our understanding of the turbulent 1930's. It demonstrates that the codes were responsible for a dramatic increase in nominal wages and prices. It was the codes which explain the coincidence of inflation and massive unemployment during the NIRA period.

0756 NH


CAN
CAPITALISM
SURVIVE?
BENJAMIN A. ROGGE

OXFORD

International Economics
The Theory of Policy

GERALD M. MEIER, Stanford University. Superbly organized, this concise text offers a systematic presentation of the fundamental normative principles that underlie commercial policy, international payments policy, and international development policy. In the tradition of political economy, the book returns to the first principles of international economic policy and focuses on the policy implications of international trade theory, international monetary theory, and international development theory strategies. "The best book for undergraduate courses in international economics."—S. Ganti, Hamilton College

1980 400 pp.; charts, tables $16.95

World Development Report, 1980

THE WORLD BANK. This year's *Report*, the third in an annual series, examines two major challenges developing countries now face: the attempt to continue their social and economic progress in an increasingly less supportive international environment; and the struggle to relieve the plight of 800 million people living in absolute poverty. In addition to illustrating the capital needs of developing countries, the *Report* includes economic projections for their growth through the end of the century.

1980 176 pp.; tables, charts cloth $13.50
paper $5.75

Money in International Exchange
The Convertible Currency System

RONALD I. McKINNON, Stanford University. "A superb treatise on the payments mechanism of international trade....As an up-to-date description and analysis of international finance, the work is unsurpassed."—*Journal of Money, Credit, and Banking.* "McKinnon has provided an extremely good 'alternative' to mainstream textbooks in the international field....its greatest strength is presenting in a straightforward, clear, and perceptive manner the activities and motives of the participants in the international marketplace."—*Economic Forum*

1979 320 pp.; tables, charts cloth $13.95
paper $7.95

Introduction to Normative Economics

E.J. MISHAN, City University, London. A lucid, well organized treatment of the foundations of prescriptive economics, this text covers all the main developments and controversies in the literature relating to welfare criteria and allocation economics over the last fifty years. Drawing a clear distinction between economic and political criteria, the author appraises the range of prescriptive propositions and reveals the limitations of welfare economics as an instrument of social betterment.

1981 576 pp.; 141 illus. paper $14.95

Information and Coordination
Essays in Macroeconomic Theory

AXEL LEIJONHUFVUD, University of California, Los Angeles. A collection of twelve papers on macroeconomics, monetary theory, and their historical development, this book provides further reflections on Keynes, the Keynesians, and the Classics while amplifying the author's own position on the main issues of unemployment and inflation theory. Several of these papers have not been readily available and two are completely new. Of particular interest is "The Wicksell Connection," a long essay clarifying the relationships among the major contending schools in modern macroeconomics.

March 1981 320 pp.; 12 figs. cloth $15.95
paper $9.95

Classical and Neoclassical Theories of General Equilibrium

Historical Origins and Mathematical Models

VIVIAN WALSH, The New School for Social Research, and HARVEY GRAM, Queens College and The Graduate School, City University of New York. "A quite remarkable book which deserves to have wide readership and, indeed, I think, to become the standard text in its area. It is lucid, scholarly, witty, fair and balanced."—G.C. Harcourt, University of Toronto

1980 448 pp.; 83 figs. $18.95

Prices and publication date are subject to change.

Oxford University Press 200 Madison Avenue New York, New York 10016

WORKSHOP 1981

The Joint Council on Economic Education

Presents Another in its Series of Summer Workshops

A week-long session—modeled on the extremely successful ones held at Indiana University in 1979 and the University of Wisconsin-Madison in 1980—will again be given for key members of economics faculties. It will take place from May 31 through June 5, 1981 at the University of North Carolina at Chapel Hill, and will be conducted by Professor Michael J. Salemi of North Carolina. Assisting staff will be drawn from the economics faculties of other institutions such as Duke, Harvard, Indiana, Minnesota, and Wisconsin. General funding and most expenses of participants will be met by a grant from the Lilly Foundation, Incorporated.

The workshop will be based on the *Resource Manual for Teacher Training Programs in Economics,* edited by Phillip Saunders, Arthur L. Welsh, and W. Lee Hansen—a manual for teaching college economics courses effectively. This unique 438-page guide is designed to be used in training programs for graduate students in economics as well as by faculty members working on their own. Although topics from the typical introductory course are used in the demonstrations, the techniques can be employed in more advanced courses. The manual is a cooperative effort of the American Economic Association's Committee on Economic Education and the Joint Council on Economic Education. Development and pilot testing were supported by a grant from the Alfred P. Sloan Foundation.

Those interested in attending the workshop should write as soon as possible to:

Dr. Arthur L. Welsh
Joint Council on Economic Education
1212 Avenue of the Americas
New York, NY 10036

CAMBRIDGE

The Economic Theory of Social Institutions

Andrew Schotter

An innovative analysis of the creation, evolution, and function of economic and social institutions using game theory. Schotter illustrates his analysis by describing the organic evolution of institutions such as the conventions of war, the use of money, property rights, and oligopolistic pricing conventions. Linking his work with the ideas of the philosophers Rawls, Novick, and Lewis, he regards institutions as regularities in the behavior of social agents—regularities which the agents tacitly create to solve a wide variety of recurrent problems. Schotter's final chapter explores the applicability of his new theory by relating it to previous work on the theory of teams, hierarchies, and non-maximizing decision theory and by using it to provide a new approach to a variety of questions both within and outside economics. $29.50

Energy Economics

Growth, Resources and Policies

Richard Eden, Michael Posner, Richard Bending, Edmund Crouch and Joe Stanislaw

A group of energy experts—economists and scientists—authoritatively surveys the present energy situation. Their integrated discussion covers the relevant physics of each major energy source (oil, gas, nuclear, and solar), present and future supply positions, and cost and pricing structure. Also examined are the connection between economics and growth, the market for energy, the effects of uncertainty on that market, and the way it is brought into balance. $34.95

The FTC Since 1970

Economic Regulation and Consumer Welfare

Kenneth W. Clarkson and Timothy J. Muris, Editors

The first comprehensive analysis of the "revitalized" FTC. Fifteen original articles discuss how the Federal Trade Commission is influenced by its institutional setting, revealing that it has broad freedom in its actions. They critically examine substantive issues characterizing FTC actions since 1970 and conclude that the FTC often causes harm to consumers as a class. $35.00

Cambridge Economic Handbooks

Theory of International Trade

A Dual, General Equilibrium Approach

Avinash Dixit and Victor D. Norman

Emphasizing that a trading equilibrium is general rather than partial and is often best modeled using dual or envelope functions, the authors give unified treatment of comparative statics and welfare and shed new light on the factor-price equalization issue. $34.50/$13.95

Economic Theory and Exhaustible Resources

P. S. Dasgupta and G. M. Heal

The authors' authoritative presentation is concerned both with the descriptive modeling of how natural resources will be used under different economic regimes and with a normative analysis of optimal use policies. $42.50/$16.95

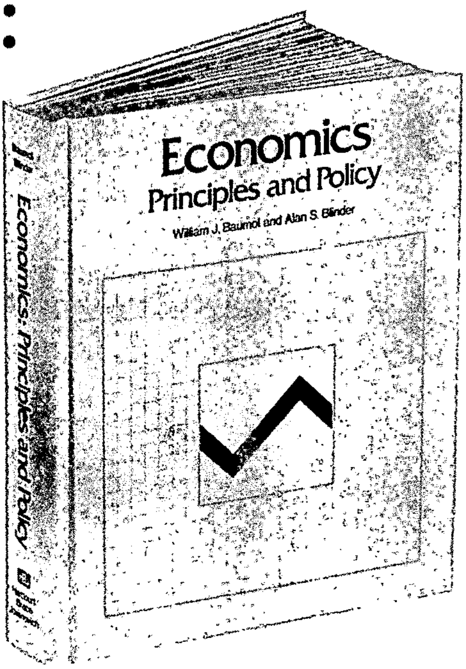
The Investment Decision of Firms

S. J. Nickell

A careful exposition of the relevant theory and a survey of the empirical research into investment behavior. Nickell makes an original contribution to synthesis and organization. $42.50/$13.95

all prices subject to change

IN DEMAND:


Economics
Principles and Policy
William J. Baumol and Alan S. Blinder
Economics: Principles and Policy

WILLIAM S. BAUMOL
New York University
and Princeton University

ALAN S. BLINDER
Princeton University

Here are some reasons why:

"Superb! [This is the] closest thing on the market to the **ideal principles text.** We are adopting it."

—M.E. McMAHON, Southwestern at Memphis

"This is an excellent book ... that deftly **integrates economic principles and applications** without contrivance and a loss of integrity.... the book offers an excellent treatment of the crucially important topics so widely discussed by reasonably educated people in today's society."

—PATRICK O'SULLIVAN, State University of New York College at Old Westbury

"An excellent—and intelligent—text ... I am most impressed with two elements: its **public-policy orientation** (which should appeal to students) and its explicit **avoidance of the encyclopedic syndrome** (which always frustrates instructors)."

—DOUGLAS E. HOUGH, Roosevelt University

"... a very well-written book that serves extremely well for both conservative and liberal instructors. It is **a dream book for students,** as it leads them to grasp the central ideas with minimum technological complications."

—S.K. KIM, California State University, Los Angeles

"[The book] takes **a fresh look at some of the real problems of society.** The chapter on the Keynesian-monetarist debate is especially crucial."

—D. LARRY DAVIS, Indiana University, Gary

862 pages

Accompanied by a **Study Guide** and an **Instructor's Manual;** a Test Bank and Transparency Masters are available on request to instructors who adopt the textbook.

HBJ

HARCOURT BRACE JOVANOVICH, INC.
New York • San Diego • Chicago • San Francisco • Atlanta

New books from the World Bank

published by Johns Hopkins

Urban and Spatial Development in Mexico
Ian Scott
Ian Scott traces the evolution of Mexico's urban system from colonial times and evaluates the effects of the country's unusually rapid urbanization. The book discusses Mexico's problems in providing its urban population with jobs, shelter, public services, and mass transportation, as well as the broader concerns of national centralization, rural-urban integration, and patterns of interregional development. Scott also suggests alternative approaches to spatial policy and the instruments that might be used to implement them. His findings are relevant not only to Mexico, but to all countries facing similar problems of rural migration. $29.50 hardcover, $9.50 paperback

Agroindustrial Project Analysis
James E. Austin
Combining a systems approach with conventional financial and economic analysis, this book presents a framework for project analysis that treats agroindustries—those enterprises that process agricultural raw materials—as one component in a larger system of related stages, from seed to consumer. The principle activities of an agroindustry—Procurement, processing, and marketing—are described in detail. Appendixes supply both the basic questions that analysts need to ask in assessing project planning and the typical costs of alternative food processing technology. $16.50 hardcover, $6.50 paperback

State Manufacturing Enterprise in a Mixed Economy
The Turkish Case
Bertil Wålstedt
Offering keen insights into the role of the state in a mixed economy, Bertil Walstedt discusses alternative industrialization strategies, describes the difficulties faced by developing countries attempting to achieve competitiveness in basic industry, explores issues in the administration and management of state enterprises, and suggests methods of making the state industrial sector truly accountable to the nation. $25.00 hardcover, $9.95 paperback

Egypt
Economic Management in a Period of Transition
Khalid Ikram
The fulfillment of Egypt's considerable economic potential, according to Khalid Ikram, hinges on careful management of such assets as oil revenues, the Suez Canal, tourism, and a well-trained labor force. This study is the most detailed examination of the Egyptian economy to appear since the mid-1960s and the first to emphasize economic management and policy. It explores issues related to population, human resources, and the major productive sectors; reviews the evolution of financial resources; and describes issues of the physical infrastructure and infrastructure investment. $32.50 hardcover, $10.50 paperback

The World Rubber Economy
Structure, Changes, and Prospects
Enzo R. Grilli, Barbara Bennett Agostini, and Maria J. t'Hooft-Welvaars
Recent oil crises and a deep economic recession have rocked the industrialized world, and have forced the producers of natural rubber to make difficult decisions if they are to remain competitive with the producers of synthetic rubber. The authors of this study assess the impact and implications of recent changes in the rubber economy in light of the structures of natural and synthetic rubber economies and the factors that determine competition between the two. Also considered are issues of production planning, pricing, profitability, and demand. $6.50 paperback

New books from Resources for the Future, Inc.

published by Johns Hopkins

Timber Supply, Land Allocation, and Economic Efficiency

William F. Hyde

Increasing recreational demands on forestland, combined with environmental restrictions, make public forest management a controversial issue. However, these pressures are more apparent than real. In this book, William F. Hyde cites the inefficiency of traditional timber management methods and suggests alternative techniques that would not only increase timber production in existing woodlands, but would also release currently marginal lands for other, higher valued uses. $19.00

Trends in Energy Use in Industrial Societies

An Overview

Joy Dunkerley, Jack Alterman, and John J. Schanz, Jr.

This book investigates the causes of variations in the energy/Gross Domestic Product ratio—the amount of energy consumed relative to a given amount of output—among selected industrialized countries and over time within a particular country. The authors find that rising energy prices, changing energy supply systems, and the composition of output are among the factors which effect changes in the energy/Gross Domestic Product ratio. $8.00 paperback

Rural Development and Urban-Bound Migration in Mexico

Arthur Silvers and Pierre R. Crosson

Since 1940, Mexico's urban population has grown 5.5 to 6.0 percent per year, with migration from rural to urban areas accounting for almost 40 percent of total urban growth. *Rural Development and Urban-Bound Migration in Mexico* examines the effects of investment in agriculture as a means to slow this dangerously heavy rural-to-urban migration and to alter the distribution of migration among cities, thus offsetting the urban overcrowding, pollution, poverty, and strains on social services that have plagued Mexico in recent years. $6.95 paperback

World Mineral Trends and U.S. Supply Problems

Leonard L. Fischman, project director

Using specially compiled sets of world statistical series, including new data on world production capacity, costs, prices, and consumption, this study identifies major supply problems for seven metallic minerals: cobalt, chromium, manganese, aluminum, copper, lead, and zinc. The book departs from traditional aggregative or implicit end-use projections of mineral supply and demand to consider instead the specific forms in which the minerals serve as feedstock for U.S. industry. $15.00 paperback

Energy Development in the Southwest

Problems of Water, Fish, and Wildlife in the Upper Colorado River Basin

edited by Walter O. Spofford, Jr.

Energy Development in the Southwest assesses the energy potential in Arizona, Colorado, New Mexico, Utah, and Wyoming. The authors attempt to provide a basic understanding of the impacts of energy development and of the nature and scope of the balance that must be struck between energy exploitation and environmental protection. Particular emphasis is placed on the fish- and wildlife-related impacts in the Upper Colorado River Basin, but the approaches and methods used are applicable in making similar assessments elsewhere. $25.00 two-volume paperback

The Johns Hopkins University Press
Baltimore, Maryland 21218

Economics for the real world.

Solutions to Unemployment

DAVID C. COLANDER, University of Miami

This stimulating and highly relevant collection of readings offers 29 articles that examine various economic policy options for coping with the problem of unemployment. The book presents a succinct overview of traditional solutions, including early statements on full employment by Keynes and Lord Beveridge, as well as current articles describing supply side solutions that represent the latest economic thinking on unemployment. Well known economists represented in the book include Friedman, Solow, Samuelson, Modigliani, Feldstein, and Lerner.

Paperbound. 224 pages (probable)
Publication: March 1981

Economics of Development and Distribution

WILLIAM LOEHR, University of Denver
JOHN P. POWELSON, University of Colorado

A comprehensive textbook for economic development, this book covers all the standard topics in the course but extends the traditional approach to the subject by integrating development literature with literature from other social sciences. One prepublication reviewer says: "We have been waiting for such a reorientation of economic thought on development in a humanistic perspective for some time. It admirably gives an overview and synthesis of the several 'new directions' in development thinking that have emerged in the last five years or so, and integrates this well with what remains valid in 'the economist's perspective' on this complicated and interdisciplinary topic area."

Geoffrey Hainsworth
University of British Columbia

448 pages (probable)
Publication: February 1981

Microeconomics
Theory and Applications

FRED R. GLAHE,
University of Colorado, Boulder
DWIGHT R. LEE, Virginia Polytechnic Institute and State University, Blacksburg

This basic introductory textbook for courses in intermediate microeconomics or managerial economics surveys all the major aspects of microeconomic theory—consumer choice, production theory and the cost of production, market structure, and welfare economics and general equilibrium—and provides many extensive applications as well as numerous short, empirically based examples. A careful review and extended applications of the concepts of supply, demand, and elasticity early in the text prepare students for the extensive use of these tools later in the text.

558 pages
Just published
Student Workbook
Instructor's Manual

HBJ

HARCOURT BRACE JOVANOVICH, INC.
New York • San Diego • Chicago • San Francisco • Atlanta

NEW BOOKS FROM MIT

AN ECONOMIC ANALYSIS OF WORLD ENERGY PROBLEMS
by Richard L. Gordon
320 pp. $30.00

INTERNATIONAL ENERGY FUTURES: Petroleum Prices, Power and Payments
by Nazli Choucri
260 pp. $27.50

STRUCTURAL ANALYSIS OF DISCRETE DATA WITH ECONOMETRIC APPLICATIONS
edited by Charles F. Manski and Daniel McFadden
588 pp. $29.95

STUDIES IN BUSINESS-CYCLE THEORY
by Robert E. Lucas, Jr.
256 pp. $17.50

THE U.S. COAL INDUSTRY: The Economics of Policy Choice
by Martin B. Zimmerman
256 pp. $25.00

Announcing a new series, Regulation of Economic Activity, edited by Richard Schmalensee, Professor of Management at MIT.

FREIGHT TRANSPORT REGULATION: Equity, Efficiency, and Competition in the Rail and Trucking Industries
by Ann F. Friedlaender and Richard H. Spady
400 pp. $30.00

THE SEC AND THE PUBLIC INTEREST
by Susan M. Phillips and J. Richard Zecher
184 pp. $17.50

Available in paperback

DICTIONARY OF ACCOUNTING
by Ralph Estes
176 pp. $4.95
(hardcover, $15.00)

DISTORTION OR DEVELOPMENT? Contending Perspectives on the Multinational Corporation
by Thomas J. Biersteker
216 pp. $7.95

INTERNATIONAL TRADE: Selected Readings:
edited by Jagdish N. Bhagwati
440 pp. $9.95
(hardcover, $25.00)

MANAGERIAL APPLICATIONS OF SYSTEM DYNAMICS
edited by Edward B. Roberts
688 pp. $17.50

THE MICROELECTRONICS REVOLUTION: The Complete Guide to the New Technology and Its Impact on Society
edited and introduced by Tom Forester
608 pp. illus. $12.50
(hardcover, $25.00)

THE PROSPECTIVE CITY: Economic, Population, Energy, and Environmental Developments
edited by Arthur P. Solomon
512 pp. $9.95

THE SCIENCES OF THE ARTIFICIAL Second Edition
by Herbert A. Simon
192 pp. $4.95
(hardcover, $15.00)

The MIT Press

Massachusetts Institute of Technology
28 Carleton Street
Cambridge, MA 02142

NEW TITLES IN ECONOMICS

MEASURING THE BENEFITS OF WATER POLLUTION ABATEMENT

By DANIEL FEENBERG and EDWIN S. MILLS

A Volume in the STUDIES IN URBAN ECONOMICS Series

CHAPTER HEADINGS: Introduction. Welfare Economics and the Basis of Benefit Measurement. Economic Theory of Benefit Measurement. Measurement of Instream Water Quality Benefits. Estimating Public Goods Demands and Demand Interdependency. Measurement of Withdrawal Benefits. Empirical Studies of Instream Benefits. An Empirical Study of Withdrawal Benefits. Estimates of National Benefits of Water Pollution Abatement. Conclusions. References. Index.

1980, 208 pp., $19.50 ISBN: 0-12-250950-1

THE GENETICS OF ALTRUISM

By SCOTT A. BOORMAN and PAUL R. LEVITT

"I . . . find it scholarly and persuasive."
—Nathan Keyfitz, Harvard University

Features of interest to economists . . .

The book describes principles of mathematical modeling in evolutionary sociobiology, with special attention paid to the natural mathematical structure that is emerging in this new area. Developments covered include an axiomatic theory of kin (sib) selection based on convexity properties of fitness coefficients, evolutionary foundations for the emergence of a division of labor, biological free-rider problems, and the use of social network models to analyze reciprocal altruism and related cooperative phenomena.

CONTENTS: The Evolutionary Roots of Sociality. THE THEORY OF RECIPROCITY SELECTION. Mathematical Models for a Simple Cooperative Trait. Cascade to Takeover by the Social Trait. Dynamics of the Cascade Using the Two-Island Approximation. The Cascade Continued—Initial Conditions and Global Dynamics. THE THEORY OF KIN SELECTION. General Models for Sib and Half-Sib Selection. Axiomatization of Sib Selection Theories. Alternative Combinatorial Models and the Status of the Hamilton Theory. Models of Intergenerational Altruism. THE THEORY OF GROUP SELECTION. Analysis of Group Selection in the Levins $E=E(x)$ Formalism. Group Selection of Founder Populations. Conclusions. Glossary. References. Author Index. Subject Index. Each chapter includes notes.

1980, 448 pp., $29.50 ISBN: 0-12-115650-8

A GUARANTEED ANNUAL INCOME

EVIDENCE FROM A SOCIAL EXPERIMENT

Edited by PHILLIP K. ROBINS, ROBERT G. SPIEGELMAN, SAMUEL WEINER, and JOSEPH G. BELL

A Volume in the QUANTITATIVE STUDIES IN SOCIAL RELATIONS Series

In the past decade, four major income maintenance experiments have been conducted in the United States. Each of these experiments tested several versions of a negative income tax (NIT), a cash assistance program offered as an alternative to the existing welfare system. The experiments utilized classical design procedures in which families were assigned to either an experimental group or a control group on the basis of a preselected set of strata. The main purpose of these experiments was to determine the effects of an NIT on work behavior and family stability, although a large number of other behavioral phenomena were also studied. This volume reports early findings from the largest and most comprehensive of the experiments, the Seattle and Denver Income Maintenance Experiments. The findings were first presented at a conference in Orcas Island, Washington in May 1978, sponsored by the Department of Health, Education, and Welfare.

CONTENTS: INTRODUCTION. *M. C. Keeley et al.*, Design of the Seattle/Denver Income Maintenance Experiments and an Overview of the Results. *H. I. Halsey*, Data Validation. EXPERIMENTAL EFFECTS ON LABOR SUPPLY. *P. K. Robins*, Labor Supply Response of Family Heads and Implications for a National Program. *R. W. West*, Labor Supply Response of Youth. *P. K. Robins and R. W. West*, Labor Supply Response of Family Heads over Time. *D. Betson et al.*, Using Labor Supply Results to Simulate Welfare Reform Alternatives. *P. K. Robins*, Job Satisfaction. *C. E. Munson et al.*, Labor Supply and Childcare Arrangements of Single Mothers. EXPERIMENTAL EFFECTS ON FAMILY BEHAVIOR. *L. P. Groeneveld et al.*, Marital Dissolution and Remarriage. *P. Thoits and M. T. Hannan*, Income and Psychological Distress. *M. C. Keeley*, Demand for Children. USING NEGATIVE INCOME TAX BENEFITS. *T. R. Johnson and J. H. Pencavel*, Welfare Payments and Family Composition. *M. C. Keeley*, Migration. *A. R. Hall*, Education and Training. *R. J. Pozdena and T. R. Johnson*, Demand for Assets. *M. E. Avrin*, Utilization of Subsidized Housing. *V. Davis and A. Waksberg*, Appendix. Index. References appear at the end of each chapter.

1980, 346 pp., $28.00 ISBN: 0-12-589880-0

FROM ACADEMIC PRESS

AMERICAN INEQUALITY
A MACROECONOMIC HISTORY

By JEFFERY G. WILLIAMSON and PETER H. LINDERT

A Volume in the INSTITUTE FOR RESEARCH ON POVERTY MONOGRAPH Series

In a major, original effort at historical reconstruction, the authors of this volume have traced inequality of income and wealth in the United States from the seventeenth century to the present time. Discovering new data in hitherto unused records, and applying sophisticated mathematical techniques to the interpretation of old data, they establish clear trends toward both inequality and equality, plot their timing and duration, examine the various hypotheses suggested to explain them, and firmly reject the notion that because inequality and economic growth went hand in hand in the nineteenth century, the first was a necessary precondition to the second. Instead, they argue, complex interactions of such variables as unbalanced technological development from sector to sector, labor supply, and rapid capital accumulation are the most likely candidates as causes for trending inequality.

CONTENTS: ISSUES AND EVIDENCE. The Issues. In the Beginning: The Distribution of Wealth in Colonial America. Trends in Wealth Inequality Since 1774. Trends in Income and Earnings Inequality Since 1816. Prices, Inequality, and Absolute Living Standards. POTENTIAL EXPLANATIONS. Competing Hypotheses. Unbalanced Growth and Labor Saving. The Factor Intensity Issue: Who Buys the Services of America's Working Poor? Labor Supply Growth. QUANTITATIVE ACCOUNTING. Explaining the Rise in Earnings Inequality Before World War I. Explaining Twentieth-Century Distribution Trends. Inequality and Accumulation: A Nineteenth-Century Growth-Equity Conflict? The Sources of Inequality. Appendices. References. Index.

1980, 380 pp., $29.50   ISBN: 0-12-757160-4

THE ECONOMICS OF UNIVERSITY BEHAVIOR

By DAVID A. GARVIN

This book approaches universities from an economic perspective, regarding them as a collection of nonprofit "firms" competing with one another in various ways. The emphasis is on identifying the distinctive organizational features of universities, modeling the ways in which they allocate resources, and tracing the implications of these decision-making rules for different kinds of institutions.

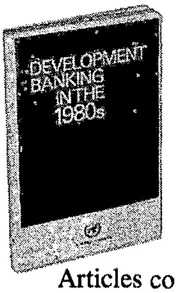
CONTENTS: INTRODUCTION. Models of University Behavior. The Economic Approach. THE HIGHER EDUCATION MARKET. THE UNIVERSITY AS AN ORGANIZATION. Financing the University. A Utility Maximizing Model. Formalizing the Model. THE INTERNAL ORGANIZATION OF UNIVERSITIES. Competition for Financial Resources. The Role of Program Cost and Prestige in Budgeting Decisions. Implications and Predictions. PATTERNS OF INSTITUTIONAL IMPROVEMENT. A Model of Prestige Improvement. Estimation Problems. Measuring Prestige Improvement. Explaining Institutional Improvement. Patterns of Upgrading Behavior. Patterns of Graduate Financial Aid. Departmental Differences. Appendices: The Cartter and Roose-Anderson Studies. Data Sources and Definitions of the Explanatory Variables. THE MARKET FOR PRESTIGE AND THE MARKET FOR DOCTORAL PROGRAMS. The Determinants of Prestige. The Market for Prestige. Empirical Evidence—Improvement by Field. The Introduction of New Doctoral Programs. Empirical Evidence—New Doctoral Programs. THE DIFFUSION OF NEW DOCTORAL PROGRAMS. The Data. The Results. Conclusion. References. Subject Index.

1980, 184 pp., $17.50   ISBN: 0-12-276550-8

CONTINUATION ORDERS authorize us to ship and bill each volume in a series, or "Advances" type publication automatically, immediately upon publication. This order will remain in effect until cancelled. Specify the volume number or title with which your order is to begin.

Send payment with order and save postage and handling.
Prices are in U.S. dollars and are subject to change without notice.

ACADEMIC PRESS, INC.
A Subsidiary of Harcourt Brace Jovanovich, Publishers
111 FIFTH AVENUE, NEW YORK, N.Y. 10003 • 24-28 OVAL ROAD, LONDON NW1 7DX



Publications from the United Nations

Development Banking in the 1980s
Selections from the Proceedings
of the UNIDO/World Bank Symposium, Zurich, June 1979

Articles collected and published in the symposium format concerning the interrelationships between government policy and industrial development banking, and the promotion of new forms of international financial co-operation.

Sales No. **E.80.II.B.3** **$13.00**

Interrelations: Resources, Environment, Population and Development
Proceedings of a United Nations Symposium held at Stockholm, 6-10 August 1979

Report on the symposium discussions, and breakdown of the agenda to four main topics: expansion of carrying capacity; development of human resources; reduction of disparities of living and adoption of appropriate life styles; and promotion of rational use of natural resources.

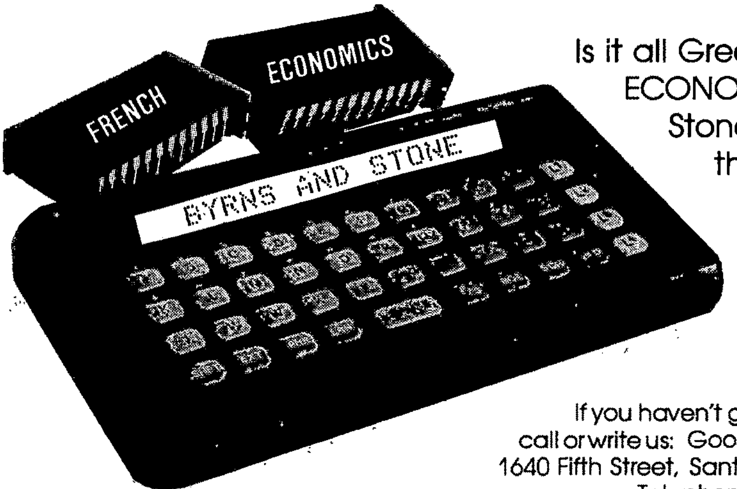
Sales No. **E.80.II.A.8** **$6.00**

United Nations Publications

Room A-3315
New York, N.Y. 10017

Palais des Nations
1211 Geneva 10, Switzerland

Now translating economics for freshmen everywhere



Is it all Greek to them? Not with ECONOMICS by Byrns and Stone. Just take a look at the analogies on page 363 or page 478. Now that's speaking their language.

If you haven't gotten your copy of this book, call or write us: Goodyear Publishing Company, 1640 Fifth Street, Santa Monica, California 90401. Telephone: 213-393-6731.

New Studies in Productivity Analysis:

New from Martinus Nijhoff, Boston

Productivity Analysis
A Range of Perspectives

Ali Dogramaci, editor
Rutgers University
Leading experts in productivity from a variety of disciplines present a range of approaches to the study of productivity in the U.S.

1981, cloth, 208 pp.
$19.95/Dfl. 45.00

Aggregate and Industry-Level Productivity Analysis

Edited by:
Ali Dogramaci and Nabil R. Adam
Rutgers University
This volume discusses methodological approaches used in aggregate and industry-level productivity studies, the movement of labor productivity measures through time, and the issues of measurement of capital and the relation of capital formation to productivity growth.

1981, cloth, 204 pp.
$25.00/Dfl. 62.50

Coming in May:

Productivity Analysis at the Organizational Level

Edited by:
Ali Dogramaci and Nabil R. Adam
Rutgers University
This volume discusses a range of productivity studies at the micro level, some productivity issues in public organizations, and current issues in productivity measurement.

1981, cloth, 192 pp.
Price: TBA

Quadrilingual Economics Dictionary

Compiled by:
Frits J. de Jong
This important reference tool is, in actuality, four dictionaries which translate economic terms from one of four base languages (English/American, French, German, and Dutch) into the other three. An indispensable reference work for all academic and corporate libraries.

1980, cloth, 685 pp.
$48.00/Dfl. 95.00

ALLES KOMT TERECHT M·N

Politicians, Legislation and the Economy
An Inquiry into the Interest-Group Theory of Government

Robert E. McCormick
University of Rochester
Robert D. Tollison
Virginia Polytechnic Institute and State University
This is a book about the behavior of politicians in supplying legislation and about how they are paid for this activity, based on the premise that they are primarily influenced by the constraints they face in the pursuit of their self-interest.

1981, cloth, pages: TBA
Price: TBA

To order, or to request a free catalog, write to:

Martinus Nijhoff, Boston

Suite 304H
190 Old Derby St.
Hingham, MA 02043

Orders from outside North America should be sent to:

Martinus Nijhoff Publishers BV
Lange Voorhout 9-11
P.O. Box 566
2501 CK The Hague
The Netherlands

Individual orders must be prepaid.
Prices are subject to change.

The American Economic Review

PAPERS AND PROCEEDINGS

OF THE

Ninety-Third Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Denver, Colorado, September 5–7, 1980

MAY 1981

THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW*, including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members six times a year, in March, May, June, September, and semimonthly in December.

Dues for 1981, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$30.00 for regular members with rank of assistant professor or lower, or with annual income of $14,400, or less;

$36.00 for regular members with rank of associate professor, or with annual income of $14,400 to $24,000;

$42.00 for regular members with rank of full professor, or with annual income above $24,000;

$15.00 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are $100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $5.00 to cover extra postage.

●Correspondance relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

Founded in 1885

Officers

President
WILLIAM H. BAUMOL
Princeton University and New York University

President-Elect
GARDNER ACKLEY
The University of Michigan

Vice Presidents
OTTO ECKSTEIN
Harvard University and Data Resources, Inc.
ALICE M. RIVLIN
Congressional Budget Office

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
ROBERT W. CLOWER
University of California-Los Angeles

Managing Editor of The Journal of Economic Literature
MOSES ABRAMOVITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
HENRY J. AARON
The Brookings Institution and the University of Maryland
ZVI GRILICHES
Harvard University
MARTIN FELDSTEIN
National Bureau of Economic Research, Inc. and Harvard University
ROBERT E. LUCAS, JR.
University of Chicago
ELIZABETH E. BAILEY
Civil Aeronautics Board
ROBERT J. GORDON
Northwestern University

EX OFFICIO Members
ROBERT M. SOLOW
Massachusetts Institute of Technology
MOSES ABRAMOVITZ
Stanford University

THE AMERICAN ECONOMIC REVIEW

VOL. 71 NO. 2 MAY 1981

PAPERS AND PROCEEDINGS

OF THE

Ninety-Third Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Denver, Colorado

September 5–7, 1980

Program Arranged by WILLIAM J. BAUMOL

Papers and Proceedings Edited by ROBERT W. CLOWER, GLENN W. HARRISON,

and WILMA ST. JOHN

CONTENTS

PAPERS

Richard T. Ely Lecture

Inventories and Macroeconomics

Applied Marxian Analyses of Stagflation

Economics of Slack Capacity

The Impact of Inflation on Labor Force Participation and the Distribution of Household Income

Resource Allocation in the Local Public Sector

Issues in Government Economic Regulation of Transportation in the 1980's

Economics of Affirmative Action

The Monetarist Contribution to Macroeconomics

The Economics of Recessions

Models for Antitrust

Testing in Economics

Distributional Issues in Public Policy

A Critical Look at Rational Expectations in Practice

Measuring Benefits from Air Pollution Control: Where Do We Stand?

Monetary Theory: Implications for Monetary Policy

Sustainability Analysis

Incomes Policy: What Did We Learn in the 1970's?

Models of Labor Market Equilibrium

The Productivity Growth Slowdown

Application of Theory to the Economics of the Arts

Inflation, Money Creation, and Government Revenue

Behavioral-Micro Approaches to Inflation

Youth Employment: New Perspectives

Freight Transport Regulation and Market Imperfections

Law and Economics in an Environment with Imperfect Information

Inflation and the Tax System

PROCEEDINGS

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

EDITOR'S INTRODUCTION

This volume contains the *Papers and Proceedings* of the ninety-third annual meeting of the American Economic Association.

The *Proceedings* record the business activities of the Association in 1980: the annual membership· meeting; the March and September meetings of the Executive Committee; reports of the Association's officers and committees.

The *Papers* constitute the greater part of the volume. They comprise seventy-six contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. The procedure governing selection of contributions for the *Papers* differs radically, of course, from that governing selection for the *American Economic Review*. About a year in advance, the Association's President-elect (in 1980 William Baumol, in 1981 Gardner Ackley), acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. The program chairman sets limits on the length of the papers at various sessions and invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing twenty-seven sessions, although a total of eighty-five sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies. There is no standard practice with regard to the publication of comments and discussions, and each program chairman must decide how to allocate available publication space between invited papers and discussions. In the present volume, unlike most earlier volumes, we are publishing nothing but invited papers.

The rules under which papers are published in the *Proceedings* are also different from those governing regular issues of the *Review*. The length of papers is strictly controlled. Except in unusual circumstances, they must be less than 4,000 (sometimes 3,000) words. Their content and range of subject matter reflect the wishes of the program chairman to investigate and expose the current state of economic research and thinking. In many cases, they are correspondingly exploratory and discursive rather than technical in character.

Although we edit the papers to improve content and style, to satisfy space requirements, and to eliminate repetition, we do not subject the papers to a refereeing process, and publication of any paper received prior to the printing deadline is virtually guaranteed if it satisfies space limitations. We would refuse to publish a paper if we concluded after reading it that it was utterly without merit, but no paper submitted this year has been rejected on those grounds. The Executive Committee has established another ground for rejection; if a paper cannot be cut to meet space limitations, we may ask the author to allow its consideration for publication in a regular issue of the *Review*, subject to the usual refereeing process, or the author may be asked to withdraw the paper and submit it elsewhere.

These practices serve a number of important purposes. The papers can be published without the long delays imposed by the refereeing process. They are short papers, covering a wide variety of subjects, and in most cases can be understood by nonspecialists. (Indeed they provide excellent text material for certain teaching purposes.) Authors have a chance to report on research to be undertaken or recently completed, to discuss topical subjects in an informal way, and to summarize longer forthcoming publications. Readers get a chance to browse among a large number of articles that are outside their major areas of interest but are not as specialized or technical as those sometimes found in field journals.

ROBERT W. CLOWER
GLENN W. HARRISON
WILMA ST. JOHN

RICHARD T. ELY LECTURE

Economics and Political Economy

By LIONEL ROBBINS*

May I begin by saying what an honor I feel it to be asked to lecture before this distinguished assembly on a foundation designed to commemorate the fame of one of the most influential economists of the earlier years of your great association. May I also say what an intense pleasure it is to be chaired by my dear friend William Baumol, an excolleague and, since our first acquaintance, the source of so much learning on my part and continuous inspiration.

Let me start by a word or two about my title which may have seemed to some of you formidably all embracing. This, let me assure you, would be a misapprehension: my target is comparatively restricted. At the beginning of my career, in my salad days, I wrote a slender essay entitled *The Nature and Significance of Economic Science*; and from time to time its contents have been the subject of criticism and discussion. I have seldom made any comment on this but I have gone on thinking. Thus, when I was invited to give this lecture, it occurred to me, with your approbation, Chairman, that, at the approaching end of my career, it might be a good opportunity to gather together some reflections on the subject of that essay and perhaps to put things in such a way as to make peace with some of my critics.

My remarks will fall broadly into four main parts. In the first—very briefly, you will be glad to hear—I shall resume my position on the definition of the subject matter of Economics. In the second I shall discuss its status as a science. In the third I shall examine the attempt to give scientific justification to the normative propositions known as Welfare Economics. And in the fourth I expound my own conception of what I now call Political Economy. In conclusion I shall try to sum up the main contentions of these somewhat discursive reflections and to point a moral as regards teaching.

I

To begin with subject matter, the conception that I argued in my book was of those aspects of behavior which, in some way or another, arise from the existence of scarcity. Now I am not at all indisposed to accept, for purposes of after-dinner conversation, Jacob Viner's wisecrack that "Economics is what economists do." But this only shifts the question one stage further: what is it that they do? What is the object of their investigations?

I hope I do not need to say much about what, in my youth, was probably the most widely used answer to this question, namely *the causes of material welfare*. Quite apart from the precise meaning of this ambiguous term, it is an easy matter to show that there is an economic aspect to the choice between the causes of material and nonmaterial welfare. And since William Baumol together with Bowen, has written a very persuasive and extensive work on *The Economics of the Performing Arts*, I think we must regard this conception as too narrow, and indeed misleading, and look elsewhere for a plausible description of the nature of our subject matter.

Much more interesting is the proposal put forward by my old friend and colleague, Fritz Hayek, to revive Archbishop Whately's proposal to rename our subject as the science of *Catallactics* (pp. 3–5), or the *Science of Exchanges*. I should certainly agree that, even where there is no market, the economic aspects of decisions and activities concern-

*London School of Economics.

ing scarce means and time can be regarded as the exchange of one state of affairs for another; and I think that this approach leads to very deep insights. But I do not think it makes sufficiently clear the conditions which lead to exchange, whether actual or implied.

But this, of course, is what the definition in terms of behavior conditioned by scarcity specifically does—scarcity being conceived as the relationship between objectives, either personal or collective, and the means of satisfying them. As you know, it first emerges in so many words in David Hume's *Treatise of Human Nature* (pp. 261–62) and it is made explicitly applicable to economic relationships in general in a famous chapter in Menger's *Grundsätze* where the limitation of goods confronted with conceivable demand is made the necessary condition of the activity of economizing. It covers exchanges and the instutional arrangements which arise in connection with this limitation.

Thus, coming back to Jacob Viner, I doubt very much that what economists do when they discuss what is, or what can be, the nature of such possibilities is not covered by this definition.

II

This brings me to the second division of my reflections. Let me say at once that I see no reason for denying to the study of the activities and institutions created by scarcity the title of science. It conforms fundamentally to our conception of science in general: that is to say the formation of hypotheses explaining and (possibly) predicting the outcome of the relationships concerned and the testing of such hypotheses by logic and by observation. This process of testing used to be called verification. But, since this way of putting things may involve an overtone of permanence and nonrefutability, it is probably better described, as Karl Popper has taught us, as a search for falsification—those hypotheses which survive the test being regarded as provisionally applicable. I am pretty sure that all the positive propositions of economics conform to this description. In this context, therefore, we may regard them as falling into the same category of knowledge as astronomy, physics, and biology—although, some may think, something of a poor relation.

But at the same time we must recognize that, within these logical criteria, the methods and problems of economic science are very substantially different from those of the so-called natural sciences. This springs from the fundamental circumstance that the subject matter is an aspect of human action and therefore must be conceived as including purpose. That is to say that our explanations must to some extent be *teleological*. This is not to argue with von Mises and some of his followers that we must regard human action, if not purely vegetative, as at all times *rational* in the sense that, given belief in the range of technical knowledge available to individuals or collections of individuals, action must be *consistent*. I confess that I have never been able to understand this contention: I should have thought that one of the main practical functions of economic science was to enable us to detect inconsistencies in plans, such as, for instance, simultaneous demands for low interest rates brought about by increases in the size of the credit base and a diminution of inflation. But, putting this conception aside, I would have thought that the contention that explanations of economic relationships must involve considerations of purposes, implicit or explicit, to be relatively noncontroversial.

Unfortunately this is not so. Influenced presumably by behaviorism in psychology, there are those who urge that in economics we must exclude any hypothesis which relies on conceptions which are not *directly observable* in the sense that they could be recorded as being perceived by the senses of an outside spectator and thus made the data of explanations of causal relationships.

I confess that I fail to see the necessity, or indeed the desirability, of the self-denying ordinance. I concede that, in the examination of simple markets, observations can be made which can be regarded as *revealing* preferences for action on the part of the persons concerned; and thus more or less determinate solutions achieved of the probable outcome. But I cannot believe that such considerations are in any way superior to those which go behind the observed dispositions to the psychological conception of *ordering* upon which the so-called subjective theory has been based in the past. And if we proceed to consider more complicated situations, I simply cannot conceive explaining to a visitor from another planet the ups and downs of a stock exchange without invoking the psychological element of expectations, not to mention error and the vagaries of fashion. According to my inadequate knowledge of physical science, I doubt whether its explanations are limited to elements which are directly observable. So long as the elements in the hypothesis are indirectly testable, they are surely scientifically admissable. Thus I ask why we, as economists, should impose on ourselves greater austerity than this?

There are, however, other differences of considerable significance between the nature of the subject matter of economics and most, if not all, natural sciences, namely, to use Paretean terms, the absence of constants both of tastes and of obstacles.

In natural science, once causal connections have been established, the quantitative relationships can usually be assumed to persist, other things being equal. It is not necessary to calculate the table of atomic weights every time particular explanations or predictions are attempted. Alas, this is not so in economics. Immense ingenuity may be devoted to establishing the conditions of demand for particular commodities; and these may sometimes help in making guesses for the future. But tastes change. A Minister of Finance would be ill advised if, in making estimates for tax purposes of the demand for cigarettes, for instance, he were to rely on computations which had been made ten years ago: he must keep himself up to date with current fashions and knowledge. The influence of the Reformation made no change in the forces of gravity. But it certainly must have changed the demand for fish on Fridays.

The same absence of persistance applies also on the side of obstacles. The human beings, whose behavior in regard to scarce goods and services is the subject of our study, are capable of learning: and learning affects conduct in various ways. Thus changes in knowledge concerning the reactions of matter in various contexts do not affect matter itself. But they may affect the possibilities of technology and therefore human action. Beyond this, knowledge concerning the results of such behavior can affect future behavior. An econometrician might discover a formula concerning the response of the Dollar Exchanges to given developments of monetary and financial policy; and if he kept it to himself, he might make a lot of money. But if he released it in the journals or the media, then it would be likely to become wrong: people would alter their financial dispositions according to the new knowledge and thus render the new knowledge erroneous.

For such reasons, quantitative prediction in economics is apt to be hazardous; much more hazardous indeed than predicting the weather. Time-series, if they have been properly collected, have status as economic history and they may serve an important rôle in testing explanations of the past. But, as a means of predicting the future, they are liable in various degrees to the vicissitudes of preferences and knowledge; and unless this is continually borne in mind, they can be seriously misleading.

This is not to say that suitably qualified propositions involving numbers should not be attempted; nor that some are less liable to error than others. Still less is it to argue that explanation of causes believed to be operating in the field of economic relationships is not a worthwhile branch of intellectual activity. There are a great many things which can be said in this connection; indeed I would say many of the most important propositions of the subject fall into the category where quantification is quite out of the question. All that is intended by the remarks I have just been making is to emphasize the differences between our subject matter and the subject matter of many natural sciences, and to draw attention to the appropriate limits which it must impose on our claims. And if, by any chance, my emphasis in this respect casts any doubt on the contention that ability to predict is the sole or necessary criterion of scientific activity, I should not feel unduly depressed. I do not think that the understanding of economic phenomena hitherto achieved, although palpably imperfect, is anything to be ashamed of.

Finally, it is important to recognize that the propositions of economics, as it has developed as a science, are positive rather than normative. They deal *inter alia* with values; but they deal with them as individual or social *facts*. The generalizations which emerge are statements of existence or possibility. They use the words *is* or *may be*, not *ought* or *should be*. There can be events or institutions having an economic aspect which we ourselves regard as ethically acceptable or unacceptable. But, in so far as the explanations of their causes or consequences are scientific, they are neutral in this respect.

It is sometimes questioned whether in the discussion of any social or economic relationships this quality of what the Germans call *Wertfreiheit* is attainable. No less an authority than Gunnar Myrdal has devoted a whole book to the argument that, explicitly or implicitly, all propositions of economic theory, all classifications of happenings having an economic aspect, must involve judgments of value. I do not agree with this position. I don't think that the proposition that, if the market is free and demand exceeds supply, prices will tend to rise, has any ethical content whatever. Nor do I concede that recognition of the consequences on investment of disparity between rates of interest and rates of return depends in the least on the political prepossessions of the economist who perceives it.

Needless to say I do not at all deny that, in the course of evolution of economics as we know it, there has been a good deal of intermixture of political and ethical discussion with the scientific discussion of fact and possibility. I shall shortly be discussing this matter further in the light of certain specific instances; and it will not appear that, *provided the logical difference between the two kinds of propositions is clearly kept in mind*, I am in the least hostile to the combination. In that youthful book of mine which evoked such fervid denunciation, I expressly denied that my position involved the view that "economists should not discuss ethical or political questions any more than the position that botany is not aesthetics means that botanists should not have views on the layout of gardens." On the contrary I went on to argue, "it is greatly to be desired that economists should have speculated long and widely on these matters." As you will see later on, my position today only involves a slight purely semantic modification of this pronouncement. I still hold that the distinction of the different kinds of propositions is inescapable and that we run the dangers of intellectual confusion on our own part and justifiable criticism from outside if we do not explictly recognize it.

III

But this brings me to the next division of my subject—the status of Welfare Economics. And since, as you may suspect, my verdict is to be somewhat adverse, let me say at once that I would yield to no one in admiration of the intentions of this development and of the ingenuity with which its analysis has often been conducted. It would not be the first time in intellectual history that dedicated efforts have led to a confusion of claims; and nothing that I am about to say must be construed as contending that these efforts were not worthwhile.

The *raison d'être* of Welfare Economics is simple. How desirable it would be if we were able to pronounce as *a matter of scientific demonstration* that such and such a policy was good or bad. Take, for instance, the removal of a protective tariff. Given information about the elasticities of demand and supply of the immediate past, we can certainly make guesses, in price and income terms, about the gains to consumers and the losses to producers of the probable outcome. There are all sorts of scientific difficulties here which I have touched upon, or hinted at, already. But the guesses, such as they are, are on an objective plane. But as soon as we move to the plane of welfare, we introduce elements which are not of that order. As in the great work of Marshall and, still more, Pigou, we are assuming that comparisons between prices and incomes before and after the event can be made a verifiable basis for comparisons between the satisfactions and dissatisfactions of the different persons involved. And that, I would urge, is not warranted by anything which is legitimately assumed by scientific economics.

Let me at once guard against a misunderstanding which has often occurred in criticisms of this position. Of course I do not deny that, in every day life, we do make comparisons between the satisfactions of different people. When the head of a family carves up a turkey, he may take account of his estimate of the satisfaction afforded to different members by different portions; and, in more serious judgments of social relationships outside the family, whenever we discuss distributional questions, we make our own estimates of the happiness afforded or the misery endured by different persons or groups of persons.

But these are *our* estimates. There is no objective measurement conceivable. Let me remind you of the fundamental issue here by comparing two situations, one of which in my judgment falls *within* scientific economics as such, and one *without*. Suppose elementary barter: A, who has a bottle of whisky, has the opportunity of exchanging it with B, who has a classical record of, say, *Fidelio*. It should be quite easy to ascertain by asking the relative valuations of the objects concerned before exchange. A relates that the classical record is worth more to him than the bottle of whisky; B contrariwise. This at no point involves interpersonal comparisons of absolute satisfaction. But now suppose that A and B fall into conversation about their respective enjoyments and A says to B, "Of course I get more satisfaction than you out of music," and B vigorously asserts the contrary. Needless to say, you and I as outsiders can form our own judgments. But these are essentially subjective, not objectively ascertainable fact. There is no available way in which we can measure and compare the satisfactions which A and B derive from music. Intelligent talk? But that may be misleading. Facial expression? That too may be deceptive. Willingness to make sacrifices of other things? But that clearly shifts the emphasis to the satisfactions derived from other things; and we are left with the ultimate difficulty of interpersonal comparisons that, as Jevons put it, "Every mind is thus inscrutable to every other mind and no common denominator of feeling seems to be possible" (p. 85).[1] Jevons' emphasis may be a bit extreme; we certainly think we know what other people are feeling, though opinions notoriously differ in different cultural settings and between different people. But it is surely incontestable where scientific proof or measurement is in question.

In this connection it is interesting to note the quite explicit agreement of Bentham to the proposition I am arguing. Among the Bentham papers, there is a passage, cited by Eli Halévy, which makes it very clear that the author of the most rigid exponent of the so-called Felicific Calculus was under no delusion that interpersonal comparability was anything but a convention—a convention, it is true, which he regarded as essential to practical reasoning. "'Tis in vain," he said, "to talk of adding quantities which after the addition will continue distinct as they were before, one man's happiness will never be another man's happiness; a gain to one man is no gain to another: you might as well pretend to add twenty apples to twenty pears. Which after you had done it would not be forty of any one thing but twenty of each as there was before" (pp. 495–96).

Now recognition of this difficulty led Pareto to the suggestion that we could only say that a community was better off if, all tastes remaining constant, a change took place which improved the position of one individual or group of individuals without making any of the rest worse off.[2] Personally I can't see anything much wrong in this from a conversational point of view. But it is clearly a judgment of value.[3] If the remaining groups regard their position rela-

[1] See also PhilipWicksteed, p. 68.

[2] See Vilfredo Pareto, pp. 617–18.

[3] This aspect of the fundamental Pareto proposition is well emphasized by Charles Rowley and Alan Peacock in their excellent book, pp. 7–25.

tively, they may well argue that the spectacle of such improvement elsewhere is a detriment to their satisfaction. This is not a niggling point: a relative improvement in the position of certain groups *pari passu* with an absolute improvement in the position of the rest of the community has often been a feature of economic history; and we know that this has not been regarded by all as either ethically or politically desirable.

An extension of the Pareto criterion which appeared first in the English literature in Jacob Viner's discussion of the effects of tariff changes in his *Studies in the Theory of International Trade* (pp. 533–34), but which owes its vast repute to its rediscovery by Lord Kaldor and Sir John Hicks, is the so-called *Compensation Principle*. According to this principle, we can still say that a community is better off, despite the fact of a change involving gains for one person or group and losses for others, if out of the gains it would be possible to compensate the losses and still leave a benefit for the gainer or gainers.

Now it is obvious that, in order that such a statement can be made, it is necessary that the compensation should actually be paid. The fact that such compensation is *conceivable* is not sufficient: if it is not *actual*, the fundamental Paretean condition is violated that while the position of one person or group is improved, the position of all others is unchanged. All that we can do if compensation is not made, is to point to the change in the positions of the gainers and the losers which at once must raise distributional considerations quite obviously involving further, and more obvious, judgments of value than are implied in the original Paretean conception!

But supposing compensation is supposed to be paid, it is still germane to point out that the practical use of such judgments which it is legitimate to make on this basis, is incomparably less than the claims originally made for Welfare Economics with capital letters. I am not blind to the negative light which the Paretean criterion must throw on the omission of externalities, positive and negative, and the problems to which they give rise; for instance, the desirability of appropriate fiscal incentives or disincentives. But I am clear that the inclusion of such factors must, in most cases, necessitate assumptions involving comparisons and contrasts of individual experiences. Still more is this true of any consideration of distributional questions. I am not against such discussions. As I shall shortly disclose I am emphatically in favor of them, in the hands of qualified persons and under appropriate labels. But with the best will in the world, I cannot help thinking that John Chipman and John Moore are right in their verdict that what they call the New Welfare Economics in an article of that name, has broken down in the strictly scientific sense and left us with the fundamental implications of the passage in Jevons which I have already quoted, namely that all recommendations of policy involve judgments of value.

V

But this brings me to my last main division. Ought we to be afraid of such assumptions? Clearly there is much to be said against such austerity, at any rate from the point of view of our usefulness to society. Politics are much too important to be left to the politicians—Adam Smith's crafty and insidious animals—and, as was the intention of my original pronouncements on this subject, if they are aware of what they are doing and do not claim scientific authority for conclusions which clearly go beyond science, there is much to be said for the practitioners of scientific economics discussing such questions of policy. They may not agree on the extra-scientific elements in their arguments. But, provided the distinction is observed, there is everything to be said for the discussions of policy to be conducted by those who are aware of the objective implications of the values on which policy rests.

But manifold problems arise even here. Let us assume for a moment the explicit adoption as a postulate of Bentham's felicific calculus, namely interpersonal comparability, each subject to be treated as equally capable of satisfaction, and use that as a basis for recommendation.

Now I make no comment on the substance of this postulate. I personally do not judge that, in any scientific sense, people are necessarily equally capable of satisfaction—whatever that may mean. I readily agree that personal entitlement in equal situations to equal treatment by law is desirable; and I would go beyond that in saying that, in personal relationships, the treatment of one's fellows on a basis of equality answers my criterion of civilized behavior. But when we come to the kind of problems with which economists interested in policy are concerned, matters become more difficult.

Let us take, for instance, the problem of direct taxation. As Edgeworth showed—without, however, recommending the conclusion— the felicific calculus, applied simply to this problem of achieving the minimum aggregate sacrifice, would involve complete equality of income. But even the most hopelessly naive would hesitate to adopt this as a practical maxim of policy. Quite apart from the tangle of administrative problems of sorting out what should be regarded as equality of income in different circumstances, there arises the quite fundamental problem of incentive—should unequal contributions receive equal remuneration? I do not think we need go further than the experience of communist states to discover that so crude an application of the idea of equal capacity for satisfaction and equal rates of diminishing marginal utility of income is really not at all helpful.

Again let me revert to the example already mentioned when I was discussing Welfare Economics—the principle of compensation for improvements involving losses elsewhere. As we have seen, it must be agreed by all exponents of this principle that in order to satisfy the fundamental Paretean criterion, it is necessary that compensation should actually be paid. But very little reflection is needed to raise doubt whether this is a sensible principle. If an improvement has been made which damages the interest of producers whose output has previously been in greater demand, is it now desirable to make payments which may have the effect of preventing movement out of the group affected? Again the problem proves to be more complicated: the solution is not to be found by a simple formula. A dynamic society needs mobility. Or does it? Is compensation to be contingent on acceptance of direction of labor? Or is that an infringement of others' rights?

And so I could go on. The burden of my remarks at this point is that formulae based on the assumptions of either the old or the new Welfare Economics are unlikely to be helpful and may well miss the main point entirely. They give at once the impression of precise guidance and yet they leave out important relevant criteria. As I have urged elsewhere, they are to be regarded as a draughty halfway house. The name conveys an impression of value-free theory which it should be just our intention to avoid.

Fortunately the evolution of terminology in this sphere provides a method of eliminating such confusion. As I said earlier on, in its beginning the label Political Economy covered a *mélange* of objective analysis and applications involving value judgments. The first three books of the *Wealth of Nations* are chiefly devoted to analysis of the market economy and its vicissitudes through history; that is to say generalized *description*. The fourth and fifth are devoted to alternative systems of policy and the functions of the state: that is to say generalized *prescription*. And until even Jevons—and after—both subjects were included under the same label, although surely the difference between the title of J. S. Mill's essay "*On the Definition of Political Economy and on the Method of Investigation Proper to it*" and the title of his *Principles of Political Economy with some of their Application to Social Philosophy* indicates a clear perception of the difference. In the last hundred years, however, beginning conspicuously, perhaps, with Alfred and Mary Marshall's *Economics of Industry* (1879), we have come to describe the generalized description as *Economics* or *Economic Science*; and the label *Political Economy*, as implying judgments of value of which we do not wish to be accused, has tended to drop out of use.

My suggestion here, as in the Introduction to my *Political Economy*: *Past and Present*, is that its use should be revived as now covering that part of our sphere of interest which essentially involves judgments of value. Political Economy, thus conceived, is quite unashamedly concerned with the assumptions of policy and the results flowing from them. I may say that this is not (*repeat not*) a recent habit of mine. In the Preface to my *Economic Planning and International Order*, published in 1937, I describe it as "essentially an essay in what may be called Political Economy as distinct from Economics in the stricter sense of the word. It depends upon the technical apparatus of analytical Economics; but it applies this apparatus to the examination of schemes for the realization of aims whose formulation lies outside Economics: and it does not abstain from appeal to the probabilities of political practice when such an appeal has seemed relevant."

It should be clear then that Political Economy in this sense involves all the modes of analysis and explicit or implicit judgments of value which are usually involved when economists discuss assessments of benefits and the reverse or recommendations for policy. In particular it deserves to be noted that the whole business of choosing index numbers falls into this conception; and surely few improvements in procedure are more desirable than recognition of this fact. But, in general, the overt recognition of the extent to which the multiplicity of proximate criteria guiding considerations of policy involve judgments of value must be wholly beneficial.

The question therefore arises what should be the ultimate values guiding us in this field. The answer must necessarily be debatable: there is no agreement yet on the ultimate desiderata of the good society: consider for example the variety of opinions regarding the desirability of growth. Speaking personally, I see no objection to regarding utility in a very wide and non-quantitative sense as one of the principle criteria. As an illustration I would cite Hume's famous discussion of the circumstances in which the institution of property is, or is not, justified, in his *Enquiry Concerning the Principles of Morals*. And since I have earlier quoted Bentham's recognition of interpersonal comparisons as a *postulate* rather than a scientific possibility, I would like to say here that, in practice, his so-called felicific calculus, far from making quantitative estimates, was actually employed in Hume's sense—a matter of judging the arrangements of society *as a going concern* according as, in a broad way, they were likely to increase pleasure or diminish pain.

Thus for instance if I were today to respond to Roy Harrod's challenge how to judge the repeal of the Corn Laws, I should not attempt to justify it in terms of the gain of utility at the expense of the producers. I should not know how to do this without comparisons which, to put it mildly, would be highly conjectural. I should base my vindication on the general utility of the extension of markets and the resulting enlargement of liberty of choice. And, as I imagine Hume would have done, I should allow for specific exceptions and stipulate conditions of tolerable and intolerable rates of change.

But, even interpreted in Hume's sense, utility is not enough—at any rate to my way of thinking. There might be utility in the broad sense in the working of the institutions of a well-run slave state, and yet the assumptions behind my Political Economy would reject them. And this would not be for the reason that the attribution of utility to such institutions was wrong—though I suspect that empirically it would be. It would be for the far deeper reason—or principle you may call it—that *acts which are not free are not acts which are capable of having value in the ethical sense.* We do not regard the movement of a herd of cattle as falling under ethical categories any more than the heat of the sun or the furious winter's rages. Only where conditions of human freedom, in some sense or other, are present do such judgments have meaning.

But the conception of liberty itself involves complications. Liberty is not anarchy. It is not a free for all—often as it is said to be by those who hate it. The idea of individual liberty does not involve liberty to curtail other people's liberty. That is why the necessity of a framework of law and an apparatus of enforcement is an essential part of the conception of a free society. But this is no facile criterion. In this connection I would cite the work of the great man after whom this lecture is named. No candid reader of Richard Ely's famous *Property and Contract* can come away believing that the sections of the law with which it deals are capable of being inscribed on two tablets of stone, or that the weighing of considerations of utility, in the sense in which I have defined it, and of the claims of liberty as the essential of conduct coming under the categories of ethics, is an easy matter.

Thus both as regards utility and liberty we are eventually involved in questions relating to the coercive powers of government and the basis of consent. I have no doubt that in the discussion of such problems considerations of Political Economy are relevant. Consider, for instance, the whole range of Adam Smith's third function of the state: "the duty of erecting and maintaining certain public works and certain public institutions which it can never be for the interest of any individual, or small number of individuals, to erect and maintain, because the profit would never repay the expense to any individual or small number of individuals, though it may frequently do much more than repay it to a great society" (pp. 687–88). Indeed, without in any way subscribing to the so-called Marxian theory of politics, I suspect that such considerations must play a very large part in any articulate theory of the state, its evolution and its activities. I welcome the growing recognition of the duty of political economists to extend their systematic investigations in this sphere; and thus, in my conception at any rate, we have quite enough on our plate in this connection to occupy us for many generations to come.

VI

Let me sum up the main points of these discursive reflections. As regards the subject matter of Economic Science, I adhere to its description in terms of behavior conditioned by scarcity. As regards its status as a science, I see no reason to deny its susceptibility to the usual logical requirements of a science, though I have emphasized the peculiar nature of its subject as concerned with conscious beings capable of choice and learning. I see no reason why we should be terrified into thinking that such analysis necessarily involves ideological bias. But beyond that, in the application of Economic Science to problems of policy, I urge that we must acknowledge the introduction of assumptions of value essentially incapable of scientific proof. For this reason, while not denying the value of some thought going under that name, I have urged that the claims of Welfare Economics to be scientific are highly dubious; and I go on to argue the lack of realism which is involved by some of the inferences which may be drawn from its assumptions. Instead I recommend what I call Political Economy which, at each relevant point, declares all relevant nonscientific assumptions; and I furnish some indications of the leading criteria and fields of speculation which should underlie this branch of intellectual activity.

One final word concerning the implications of this conception of the task of Political Economy. I venture to suggest that, as teachers of the subject, our instructions will be more fruitful if, side by side, they run parallel with suitable courses in Politics and History—Politics because it deals systematically with philosophical and constitutional matters which as regards Political Economy only arise incidentally; History, because while it certainly does not lay down laws by which we can foretell the future, it does give a feeling for the possibilities of action which confining our attention to the present certainly fails to convey. I fancy that such exhortations are more at home in my own country where excessive specialization in the first-degree stage, productive of one-eyed monsters, is too frequently the order of the day. But the general principle seems to me to be sound.

REFERENCES

J. S. Chipman and J. C. Moore, "The New Welfare Economics, 1939 – 1974," *Int. Econ. Rev.*, Oct. 1978, *19*, 547–84.

F. Y. Edgeworth, *Papers Relating to Political*

Inventories and the Structure of Macro Models

By ALAN S. BLINDER*

The message of this paper can be summed up in two words: *inventories matter*. They matter empirically, in the sense that inventory developments are of major importance in the propagation of business cycles; and they matter theoretically, in the sense that recognition of their existence changes the structure of a variety of theoretical macro models in some fairly important ways. This paper is mainly about the implications of inventories for the structure of theoretical macro models, but I begin by demonstrating the empirical importance of inventories in business fluctuations.

I. The Importance of Inventories in Business Cycles

Inventory investment is a tiny component of *GNP*, averaging only about 1 percent of the total, but its importance in business fluctuations is totally out of proportion to its size. As Table 1 shows, inventory investment typically accounts for about 70 percent of the peak-to-trough decline in real *GNP* during recessions.

Of course, recessions are rather special episodes. To get a broader perspective, note that real *GNP* (Y_t) is the sum of real final sales (X_t) and real inventory investment (ΔN_t, where N_t is the stock of inventories). After detrending each series and first differencing, we have $\Delta y_t = \Delta x_t + \Delta^2 n_t$, where lower case letters denote deviations from trend. It follows that the variance of changes in the deviations of *GNP* from trend can be decomposed as follows:

$$\underset{90.4}{Var(\Delta y)} = \underset{59.1}{Var(\Delta x)} + \underset{33.4}{Var(\Delta^2 n)} + \underset{-1.8}{2\,cov(\Delta x, \Delta^2 n)}$$

where the empirical magnitudes for the United States during 1959:1–1979:4 appear below each symbol. Changes in inventory investment account for 37 percent of the variance of changes in *GNP*. The importance of inventory fluctuations is not limited to cyclical downturns.

What types of inventories predominate in these inventory fluctuations? For the period 1959–76, unpublished quarterly data from the Bureau of Economic Analysis enable us to break down real nonfarm inventories into the six components listed in Table 2. The table shows that the predominant type of inventories accounting for variation in $\Delta^2 n$ are retail inventories, followed by manufacturers' inventories of raw materials and wholesalers inventories. Neither manufacturers' finished goods nor works in progress contribute much to the variance of $\Delta^2 n$. Note also in Table 2 that the correlations between Δx and the components of $\Delta^2 n$ are all pretty meager.

II. Microfoundations

The standard theory of the firm is based on nonstorable output. When output is storable, however, firms have an additional degree of freedom: they are able to make current production Y_t differ from current sales X_t, and often will find it advisable to do so. They may use inventories of finished goods to speculate on future price movements or to absorb short-run shocks to de-

*Princeton University and National Bureau of Economic Research. I thank the National Science Foundation for research support and David Ashmore for research assistance. A slightly longer version is available as National Bureau of Economic Research Working Paper No. 515.

TABLE 1—CHANGES IN *GNP* AND IN INVENTORY INVESTMENT IN THE POSTWAR RECESSIONS

Dates of Contraction		Decline in Real *GNP*[a]	Decline in Inventory Investment[a]	Col. (3) as a Percentage of Col. (2)
Peak	Trough			
(1)		(2)	(3)	(4)
1948:4	1949:4	$ 6.7	$13.0	194
1953:2	1954:2	20.6	10.2	50
1957:3	1958:1	22.2	10.5	47
1960:1	1960:4	8.8	10.5	119
1969:3	1970:4	12.0	10.1	84
1973:4	1975:1	71.0	44.8	63

Source: *The National Income and Product Accounts of the United States*, 1929-74, and *Survey of Current Business*.

[a] In billions of 1972 dollars.

TABLE 2—DECOMPOSITION OF THE VARIANCE OF $\Delta^2 n$, 1959:1–1976:4

Inventory Component	Variance	Percent of Total Variance	Correlation with Δx
Total Inventories ($\Delta^2 n$)	40.40	100	−.02
Manufacturer's Inventories:			
Finished Goods	2.15	5.3	−.05
Works in Progress	2.45	6.1	+.25
Materials and Supplies	5.20	12.9	−.09
Wholesale Inventories	4.77	11.8	−.07
Retail Inventories	14.18	35.1	+.18
Other[a]	4.13	10.2	−.11
All Covariance Terms	7.27	18.0	−−

[a] Includes other nonfarm inventories plus statistical discrepancies that arise because the disaggregated components of manufacturers' inventories have not been revised while the total has been revised.

mand; they may use inventories of raw materials to hedge against future price increases. Inventory holdings may be used to spur demand (by reducing delivery lags) or to reduce production costs (through improved scheduling).[1]

The first point is fairly obvious: the existence of inventories requires a new concept of market equilibrium. Since it may well be optimal for firms to set $Y_t \neq X_t$, there is no reason to think that "equilibrium" means that the market "clears" in the usual sense ($X_t = Y_t$). Instead, an appropriate definition of equilibrium seems to be a situation in which the quantity that suppliers desire to sell equals the quantity that customers desire to buy. Note that Y_t is not even involved in this definition: it can, in principle, be anything.

The second point is that profit maximization probably dictates that the beginning-of-period inventory stock N_t affects firms' decisions. Specifically, I wish to argue that output, sales, and inventory carryover depend on N_t as follows:

$$(1a) \qquad Y_t = Y(N_t) \qquad -1 < Y'(\cdot) < 0$$

$$(1b) \qquad X_t = X(N_t) \qquad 0 < X'(\cdot) < 1$$

$$(1c) \qquad N_{t+1} = F(N_t) \qquad 0 < F'(\cdot) < 1$$

These equations have several obvious macro implications, and one that is not so

[1] The theoretical analysis of this paper pertains exclusively to inventories of finished goods. Different analyses would be necessary for inventories of inputs and works in progress.

obvious. Equation (1a) implies that models of aggregate supply—such as the celebrated Lucas supply function— should allow production to depend on inventory stocks. Equation (1b) implies (via the law of demand), that higher inventories lead to lower prices. Taken together, (1a) and (1b) imply that *GNP* and final sales may sometimes exhibit rather divergent behavior during short-run business fluctuations. Equation (1c) suggests that inventory investment equations should have a "partial adjustment" form, even in the absence of explicit costs of changing either production or inventory levels.

While it is possible to derive results like (1) rigorously in the context of specific micro models, I prefer to rely on an intuitive argument because it suggests that the equations are much more general than any specific model. The basic idea can be explained with the aid of Figure 1, which depicts (as point C) the equilibrium of the textbook firm with nonstorable output: optimal production (=sales) is determined by equating marginal revenue (MR) to marginal cost (MC). But when output is storable, the firm must operate simultaneously on *two* margins. To decide how many inputs to turn into inventories, it equates MC to the shadow value of inventories, which I call λ (point B). To decide how many inventories to withdraw for sale, it equates λ to MR (point A). Obviously, these separate decisions need not lead to $X=Y$.

The implications of equations (1) follow from Figure 1. So long as MC is an increasing function of Y and λ is a decreasing function of N,[2] it is clear that Y is a decreasing function of N, in accord with (1a). Similarly, so long as MR decreases with X, it is clear that X is an increasing function of N, in accord with (1b). Rising marginal costs also imply that it is optimal to rectify inventory imbalances gradually, in accord with (1c).

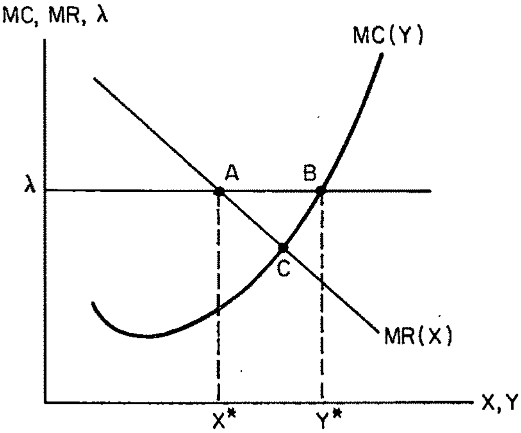


FIGURE 1

I now come to my third point, which is the nonobvious implication of (1): as compared to a world with nonstorable output, prices become "sticky" when output is storable.[3] The reason, of course, is the buffer-stock role of inventories. When there is a temporary surge in demand, the necessary price increase is moderated by the fact that firms disgorge inventories. In terms of Figure 1, the shadow value of inventories λ should be relatively insensitive to transitory shifts in demand (or cost) because it depends on *all* future demand and cost functions. Consequently, a shift in the MR schedule induces a larger sales response (a smaller price response) when output is storable than when it is not.

I close this section on microfoundations with a (loosely stated) general proposition about price rigidity that is prompted by these remarks:[4] Prices are more "rigid," that is, respond less to demand shocks, when the costs of varying inventory levels are lower and when demand shocks are less persistent.

The following sections use these microfoundations to develop the implications of inventories for the specification and logical structure of a variety of macro models.

[2] This negative relationship between inventories and their shadow value is slightly more subtle than might be expected. Under perfect competition, the shadow value of inventories can never diverge from the market price because firms can always sell *or buy* unlimited quantities at the going price. I therefore assume differentiated products with downward-sloping demand curves. For further details, see my 1978 paper.

[3] This implication is brought out by Louis Phlips, P. Reagan, and Y. Amihud and H. Mendelson.

[4] A more precise statement and a proof can be found in my 1980b paper.

III. Inventories and Old-Fashioned Keynesian Models

By including the stock of inventories in standard, old-fashioned Keynesian models, we simultaneously rid them of a serious logical flaw and of what is sometimes considered their most distressing empirical prediction—that real wages move countercyclically.

The logical flaw is quite general, but I will illustrate it with the simplest possible fixed-price model. The question is: what forces drive the economy toward an equilibrium where $Y=C+I+G$? A perfectly coherent answer is provided in most elementary textbooks. If Y, for example, exceeds $C+I+G$, inventories begin piling up, and this inventory disequilibrium signals firms to cut production. The problem is that this intuitive answer tends to get lost when models are formalized and mathematized. For example, a typical adjustment mechanism is (see, for example, Paul Samuelson, pp. 276–283):

$$\dot{Y}=\beta(X-Y),\ \beta>0$$

where X is final sales. This tacitly defines equilibrium as any state in which inventories are *constant* ($X=Y$), regardless of the *level* of inventories—in stark contradiction to the intuitive story just related. In my 1977 article, I explore this problem and suggest a resolution based on (1a) which makes the planned level of production (not the change) depend on the level of inventories (not the change).

When this basic idea is embodied in a full-fledged Keynesian model with an endogenous price level (see my 1980a paper), a number of interesting results emerge. These may be listed briefly:

1) Instead of the countercyclical behavior of real wages predicted by standard Keynesian models, search-theoretic models, and "new classical" models with rational expectations, the Keynesian model with inventories predicts that real wages move procyclically. The reason is that inventory fluctuations cause the demand curve for labor to shift along a stable labor supply function during business fluctuations.

2) The dynamic adjustment path following an increase in aggregate demand includes a period during which inflation is accelerating while output is falling. Thus inventory adjustments offer yet another instance of stagflation of the "overshooting" variety.

3) The association between inventories and output is *countercyclical* in the very short run, but predominately *procyclical* over business fluctuations.

IV. Inventories and "Disequilibrium" Models

The existence of inventories has profound implications for the recent wave of so-called (and badly misnamed) "disequilibrium" macro models. Indeed, I would go so far as to say that it robs them of much of their interest.

Among the fundamental notions of these models are the "min condition" of voluntary exchange and the concept of "spillovers" from one market to another. But the existence of inventories undermines both of these. For example, in simple disequilibrium models such as Barro-Grossman, a firm facing a sales constraint due to a non-market-clearing price sells *and produces* the minimum of its notional supply and the constraint itself. If the constraint is binding, therefore, it cuts back on production, and hence on employment. So excess supply in the good market "spills over" into the labor market.

Now suppose that output is storable, so that production and sales can diverge. A firm confronted with a short-run sales constraint may find it optimal to produce more than it can sell, adding the unsold balance to its inventories. So output is not the minimum of "notional" supply and sales. To the extent that firms provide *more* than the "min condition" dictates, any spillover of excess supply of goods into excess supply of labor is curtailed. Thus inventories provide a buffer stock—or, as Axel Leijonhufvud put it, a "corridor" that limits the applicability of standard disequilibrium analysis to instances of truly severe shocks.

Recognition of the buffer-stock role of inventories also gets rid of the most embarrassing empirical prediction of the

Barro-Grossman model. Under conditions of excess demand in the goods market, Barro-Grossman workers, unable to purchase all the goods they want, curtail their supply of labor. Thus excess demand for goods spills over into excess demand for labor. Via this mechanism, an increase in aggregate demand, starting from a position of equilibrium, will actually *reduce* output. With buffer stocks of inventories, of course, only very extreme demand shocks will render workers unable to buy the goods they want. As long as we remain in Leijonhufvud's corridor, increases in aggregate demand increase output regardless of whether the economy is initially in a state of equilibrium, of excess demand, or of excess supply (see my 1980a paper).

V. Inventories and "New Classical" Models

Recent developments in macro theory have been dominated by the new classical models. The basic ingredients of these models are continuously clearing markets, rational expectations, and some variant of the Lucas supply function:

$$(2)\qquad Y_t = K_t + \gamma(p_t - {}_{t-1}p_t) + e_t$$

where p_t is (the *log* of) the price level, ${}_{t-1}p_t$ is its expectation formulated at time $t-1$, and e_t is a white noise disturbance. Such models do not exhibit serially correlated output disturbances unless we assume some sort of adjustment costs or accelerator mechanism for the capital stock, and they imply that fully anticipated monetary policy has no real effects.

I think it fair to say that these models have not paid much attention to the fact that many outputs are storable. Consider what happens if we maintain the assumption of continuous market clearance, but replace (2) by a supply function augmented along the lines of the microfoundations suggested in Section II:[5]

$$(3)\quad Y_t = K_t + \gamma(p_t - {}_{t-1}p_t) + \lambda(N^*_{t+1} - N_t) + e_t,\ 0 < \lambda < 1$$

where N^*_{t+1} connotes the desired level of inventories. Note that (3) obeys the principal implication of (1a). Similarly, assume that in accord with (1c):

$$(4)\quad N_{t+1} - N_t = \theta(N^*_{t+1} - N_t) - \phi(p_t - {}_{t-1}p_t) + v_t,\ 0 < \theta < 1$$

In this model, it is easy to see that unanticipated price-level shocks give rise to serially correlated output disturbances: A positive price surprise of one unit initially raises output by γ and reduces inventories by ϕ. If there are no further shocks, the resulting inventory shortage will be corrected gradually according to the adjustment parameter θ; and so long as inventories remain below N^* output will remain above its full-information (natural) level.

Anticipated monetary policy will have real effects *if* desired inventories are sensitive to real interest rates *and* real interest rates are sensitive to anticipated changes in money.

In addition to these theoretical propositions, (3) and (4) have implications for empirical work on the effects of unanticipated money. According to the model, the lagged effects of unanticipated money on output that Robert Barro has found are entirely due to inventory (and unfilled orders) discrepancies caused by past unanticipated money. Empirical evidence on this implication is mixed. (See William Haraf, Steven Sheffrin, and Robert Gordon.)

Finally, consider the possibility that markets do not clear because prices are "sticky" in some sense. A well-known paper by Bennett McCallum pointed out that some types of price rigidity leave intact the characteristic prediction of new classical models that only unanticipated money has real effects. However, R. Frydman has criticized McCallum's model for foundering on the logical pitfall mentioned at the start of Section III: it fails to take account of the effects of inventories on production decisions (in accord with (1a)). Frydman shows that a more appropriate treatment of inventories (along the lines of (3)) leads to the conclusion that anticipated monetary policy has real effects when prices are sticky.

[5] The following paragraphs summarize the findings of my forthcoming article with Stanley Fischer.

REFERENCES

Y. Amihud and H. Mendelson, "Monopoly under Uncertainty: The Enigma of Price Rigidity," mimeo., Mar. 1980.

Robert J. Barro, "Unanticipated Money Growth and Unemployment in the United States," *Amer. Econ. Rev.*, Mar. 1977, *67*, 101–15.

______**and Herschel I. Grossman,** *Money, Employment and Inflation*, Cambridge University Press, Cambridge 1976.

A. S. Blinder, "A Difficulty with Keynesian Models of Aggregate Demand," in his and P. Friedman, eds., *Natural Resources, Uncertainty and General Equilibrium Systems: Essays in Honor of Rafael Lusky*, 1977.

______, "Inventories and the Demand for Labor," mimeo., Princeton Univ., Mar. 1978.

______(1980a) "Inventories in the Keynesian Macro Model," *Kyklos*, 1980.

______(1980b) "Inventories and Sticky Prices: More on the Microfoundations of Macroeconomics," mimeo., July 1980.

______**and S. Fischer,** "Inventories, Rational Expectations, and the Business Cycle," *J. Monet. Econ.*, forthcoming.

R. Frydman, "A Note on Sluggish Price Adjustments and the Effectiveness of Monetary Policy under Rational Expectations," *J. Money, Credit, Banking*, forthcoming.

R. J. Gordon, "Discussion of William Haraf, 'Tests of a Natural Rate Model with Persistent Effects of Aggregate Demand Shocks'," paper presented to the Nat. Bur. Econ. Res. Conference on Inventories, Princeton, Mar. 1980.

W. Haraf, "Tests of a Natural Rate Model with Persistent Effects of Aggregate Demand Shocks," paper presented to the Nat. Bur. Econ. Res. Conference on Inventories, Princeton, Mar. 1980.

A. Leijonhufvud, "Effective Demand Failures," *Swedish J. Econ.*, Mar. 1973, *75*, 27–48.

B. T. McCallum, "Price-level Stickiness and the Feasibility of Monetary Stabilization Policy with Rational Expectations," *J. Polit. Econ.*, June 1977, *85*, 627–634.

L. Phlips, "Intertemporal Price Discrimination and Sticky Prices," *Quart. J. Econ.*, May 1980, *92*, 525–42.

P. B. Reagan, "Inventories and Asymmetries in Price Adjustment," mimeo., MIT, Apr. 1980.

Paul A. Samuelson, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, Mass., 1947.

S. M. Sheffrin, "Inventories, Rational Expectations and Aggregate Supply: Some Estimates," paper presented to the Nat. Bur. Econ. Res. Conference on Inventories, Princeton, Mar. 1980.

Investment in Finished Goods Inventories: An Analysis of Adjustment Speeds

By LOUIS J. MACCINI AND ROBERT J. ROSSANA*

It is well known that fluctuations in inventory investment are a major source of fluctuations in gross national product. This has stimulated numerous empirical studies designed to understand the causes of changes in inventory investment. These studies include those (for example, Michael Lovell) that utilize a flexible accelerator model of inventory behavior as well as those (for example, David Belsley) that utilize the linear decision rule approach to optimal inventory holding. The latter approach, however, can be interpreted in terms of a flexible accelerator model. See J. C. R. Rowley and P. K. Trivedi for a survey of the literature.

Recently, several authors (for example, Martin Feldstein and Alan Auerbach) observed that these studies yield very implausible empirical results. The basic difficulty is that the estimates of the speed with which firms close gaps between desired and actual inventory stocks are very low. Often, the estimated speed of adjustment is less than 10 percent per quarter. As Feldstein and Auerbach stress, this is extremely implausible when even the largest swings in inventories in a given quarter amount to less than one day's production.

The purpose of this paper is to undertake an analysis of adjustment speeds for finished goods inventory investment. In our empirical work, we utilize a modified flexible accelerator model of inventory investment.[1] Like the conventional flexible accelerator, the model permits firms to close a fraction of the gap between desired and actual inventories in any period. The relevant fraction is of course the adjustment coefficient, and it may vary in principle between zero and unity. Our model differs from the conventional model in assuming that the desired stock of finished goods inventories depends on the normal levels of exogenous variables that firms must forecast to make decisions on inventory holdings. These include not only the normal level or orders, or demand, which is a standard explanatory variable in the empirical literature, but also the normal levels of real factor-input prices and real interest rates. In addition, we permit the normal levels to change relatively slowly in response to changes in past levels of orders, real factor-input prices, and real interest rates. Our objective is to investigate whether the slow adjustment speeds that have been estimated in the literature are due to the use of models which contain an incomplete menu of exogenous variables to determine desired inventories and pay inadequate attention to lags in the adjustment of normal levels of exogenous variables.

*Associate professor, The Johns Hopkins University, and economist, Federal Reserve Bank of Philadelphia, respectively. We wish to thank C. Kuduk and D. Robinson for their able research assistance, J. Carlson, C. Christ, and H. Kawai for comments, and J. Hinrichs, U.S. Department of Commerce, for kindly providing the data on deflated inventories of finished goods. The views expressed herein are ours alone and do not represent the views of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System. A longer version of this paper is available from either of us upon request.

[1] The model used here is similar to the target-adjustment model used by Feldstein and Auerbach in their empirical work. There are, however, two important differences. First, they assume that desired stocks change slowly because firms adjust slowly their target stocks in response to changes in expected sales. We assume that desired stocks change slowly because the normal variables that determine them respond slowly to changes in actual values. Further, desired stocks in our model depend not only on normal orders, or sales, but also on normal factor input prices and real interest rates. Secondly, unlike Feldstein-Auerbach who assume that the adjustment coefficient is unity, we test to see whether in fact it is unity.

I. The Model

The model underlying the behavioral relationship for inventory investment that we will use in the empirical work was developed by Maccini for the purpose of analyzing price behavior. Since an analysis of the model is presented in his earlier paper, we will merely present here an outline of the model, emphasizing the features of the model that are needed for an analysis of inventory investment.

The model envisions a firm that makes price, output, and finished goods inventory decisions. Inventories are held by the firm to satisfy buffer-stock motives. The firm makes inventory decisions by balancing at the margin the benefits from holding inventories against the costs. The benefits are essentially protection from "stock-outs," while the costs include the usual storage and interest charges. The price and output decisions serve essentially as the instruments that the firm uses to accumulate or decumulate optimally its inventory stock to bring it to desired levels.

The behavioral relationship for planned investment in finished goods inventories that emerges from the model may be written in "flexible accelerator" form:

$$(1a) \qquad \Delta lnH_t = \lambda(lnH_t^* - lnH_{t-1})$$

$$(1b) \qquad 0<\lambda\leqslant 1$$

where H_t is the firm's actual stock of finished goods inventories held at the end of period t, H_t^* is its desired or normal stock of inventories where the determination of the desired stock is made at the beginning of period t, λ is the adjustment coefficient which measures the fraction of the gap between the desired and actual stocks that is closed in a period, and Δ is the first difference operator.

The desired stock of inventories is determined by

$$(2a) \qquad lnH_t^* = \Gamma_0 + \Gamma_1 lnQ_t^{e*} + \Gamma_2 ln(W_t^{e*}/P_t^{e*}) + \Gamma_3 ln(V_t^{e*}/P_t^{e*}) + \Gamma_4 R_t^{e*}$$

$$(2b) \qquad \Gamma_1>0\,\Gamma_2<0\,\Gamma_3<0\,\Gamma_4<0$$

where Q_t^{e*} is the expected normal level of industry orders, W_t^{e*} is the expected normal level of the money wage rate, V_t^{e*} is the expected normal level of raw material prices, P_t^{e*} is the expected normal average price level prevailing in the industry, and R_t^{e*} is the expected normal real rate of interest.

The determinants of the firm's desired stock are the normal levels of variables that it must forecast to make decisions. First, the firm's desired stock should be positively related to normal industry orders. The representative firm is assumed to face demand conditions in which the new orders that it plans to receive are proportional to normal industry orders. The factor of proportionality is the firm's expected market share, which the firm is assumed to be able to control by varying its price. But, normal industry demand is assumed to be completely price inelastic (i.e., independent of the industry price level) and, consequently, normal industry orders become a determinant of H^*. (See Maccini for some evidence that this is a reasonable approximation with highly aggregative data.) In addition, the firm's desired stock should vary inversely with the normal levels of real wage rates and real raw material prices, which are components of production costs, and with real interest rates, which are a component of inventory holding costs.

To use this model in empirical work, however, two matters need to be attended to. First, (1a) is a relationship for planned inventory investment — more precisely, planned differences in the logarithm of inventories. We add an error term, ε_t, to (1a) to account for unplanned inventory accumulation. Second, the normal levels of orders, real factor-input prices, and real interest rates which are the determinants of H_t^* are of course unobservable and must therefore be related to observable variables for estimation to proceed. We assume that expectations are formed autoregressively which means that each normal variable is assumed to be a distributed lag function of past actual levels of itself.

Making use of these adjustments and substituting (2) into (1) yields

$$\text{(3a)}\quad \Delta lnH_t = \alpha_0 - \lambda lnH_{t-1} + \sum_{i=1}^{L_1} \alpha_i^1 lnQ_{t-i} + \sum_{i=1}^{L_2} \alpha_i^2 ln(W_{t-i}/P_{t-i}) + \sum_{i=1}^{L_3} \alpha_i^3 ln(V_{t-i}/P_{t-i}) + \sum_{i=1}^{L_4} \alpha_i^4 R_{t-i} + \varepsilon_t$$

$$\text{(3b)}\quad \Sigma\alpha_i^1 = \lambda\Gamma_1\Sigma\omega_i^1 > 0;\ \Sigma\alpha_i^2 = \lambda\Gamma_2\Sigma\omega_i^2 < 0$$

$$\text{(3c)}\quad \Sigma\alpha_i^3 = \lambda\Gamma_3\Sigma\omega_i^3 < 0;\ \Sigma\alpha_i^4 = \lambda\Gamma_4\Sigma\omega_i^4 < 0$$

which is our estimating equation. The $\omega_i^j \geqslant 0$ are the weights of the distributed lag that defines the jth normal variable.

II. Empirical Results

The model was tested with *U.S.* data from Total Manufacturing and the Nondurable and Durable goods components of Total Manufacturing. We used quarterly, seasonally adjusted data which covered the period 1964I–1976IV.[2]

To choose the lag length for each distributed lag in (3), we undertook a search process which selected the best-fitting combination of lag lengths as the one that generally minimized the standard error of estimate. The distributed lag was estimated freely when the length of the lag was set at four or less. When the length of the lag was set at a value greater than four, the Almon Method (with third-degree polynomials and no endpoint constraints) was used to estimate the lag distribution in order to conserve on degrees of freedom.

The empirical results are presented in Table 1. We began our analysis with a basic model in which we assumed that $\Gamma_2 = \Gamma_3 = \Gamma_4 = 0$ in (2). These restrictions reduce the model to a flexible accelerator in which the desired stock of inventories depends solely on the normal level of orders. Although this is a restrictive form of the model, it is comparable to the standard flexible accelerator model used in the literature (see Lovell) and is therefore a useful starting point.

We first fitted an equation with a relatively short distributed lag on orders, namely, four quarters. This is similar to several flexible accelerator models in the literature where short distributed lags (for example, moving averages) on sales or orders have been used. As equation (i) of each sector indicates, the model fits poorly. The estimates of the adjustment coefficient are extremely low—less than 5 percent per quarter. This, of course, is the result that Feldstein and Auerbach have persuasively criticized as being implausible. Moreover, the effect of demand (i.e., normal orders) on investment is either very weak or of the wrong sign.

We then varied the length of the lag on normal orders. The equation with the lag length that minimized the standard error of estimate is reported as equation (ii) for each sector. Lengthening the lag on orders certainly resulted in an improved fit of the model for each sector. In each case, S_e was substantially lowered, the estimate of the adjustment parameter increased, and the level of normal orders now has the correct sign and is significant. These results thus provide some support for the view that desired inventory stocks depend on the normal level of orders where the latter changes slowly. Nevertheless, problems remain. The adjustment parameter is still implausibly low, and the overall fit of the model is rather poor.

This led us to expand the basic model to include normal factor-input prices as determinants of desired inventories. Regression (iii) in each sector reports the results of adding a distributed lag of real wage rates as a measure of normal real wage rates to the basic model. We searched over various

[2] The data used include the Department of Commerce's new series on real finished goods inventories, deflated new orders, average hourly earnings excluding overtime, the wholesale price index for crude materials, the short-term business loan rate, and producer wholesale price indices for output prices. In applying (3) to aggregate data, we are following customary procedure in the literature on inventory investment by assuming that (3), which describes the behavior of an individual firm, holds in the aggregate as well.

TABLE 1—EMPIRICAL RESULTS FROM ESTIMATING EQUATION (3a)

	Parameter Estimates[a]					Summary Statistics[b]		
Sector and Equation	$\hat{\lambda}$	$\Sigma\hat{\alpha}_i^1$	$\Sigma\hat{\alpha}_i^2$	$\Sigma\hat{\alpha}_i^3$	$\Sigma\hat{\alpha}_i^4$	$\bar{R}^2$	S_e	Test for Auto-correlation
Total Manufacturing								
(i)	.042	.018	–	–	–	.148	.01034	.480
	(.022)	(.024)[4]						(.140)
(ii)	.097	.084	–	–	–	.328	.00872	.084
	(.058)	(.045)[24]						(.045)
(iii)	.388	.906	−1.185	–	–	.513	.00743	.340
	(.118)	(.214)[24]	(.288)[20]					(.218)
(iv)	.478	.405	–	−.152	–	.591	.00680	.354
	(.091)	(.076)[18]		(.032)[10]				(.189)
(v)	.737	.901	−.545	−.180		.658	.00622	.207
	(.166)	(.199)[24]	(.222)[18]	(.079)[10]	–			(.254)
(vi)	.195	.164	–	–	−.0027	.334	.00869	.574
	(.118)	(.108)[24]			(.0031)[20]			(.212)
(vii)	.780	1.163	−1.232	−.181	.018	.743	.00540	.170
	(.148)	(.225)[24]	(.337)[18]	(.133)[10]	(.012)[20]			(.311)
Nondurable Goods								
(i)	.020	−.013	–	–	–	.130	.01305	−.020
	(.040)	(.037)[4]						(.500)
(ii)	.160	.117	–	–	–	.337	.01082	.167
	(.045)	(.044)[24]						(.159)
(iii)	.500	.728	−.410	–	–	.542	.00899	.214
	(.118)	(.163)[12]	(.125)[16]					(.202)
(iv)	.186	.153	–	−.060	–	.473	.00965	−.012
	(.093)	(.106)[24]		(.103)[8]				(.190)
(v)	.978	1.929	−2.023	−.394	–	.690	.00740	.010
	(.154)	(.361)[18]	(.443)[22]	(.253)[16]				(.357)
(vi)	.224	.193	–	–	−.0008	.413	.01020	.118
	(.099)	(.103)[24]			(.0024)[20]			(.221)
(vii)	.942	2.246	−2.454	−.220	.0097	.739	.00679	.072
	(.145)	(.355)[18]	(.479)[22]	(.277)[16]	(.0091)[12]			(.356)
Durable Goods								
(i)	.048	.039	–	–	–	.215	.01205	.354
	(.017)	(.021)[4]						(.148)
(ii)	.143	.128	–	–	–	.327	.01059	.355
	(.032)	(.033)[12]						(.152)
(iii)	.262	.260	−.234	–	–	.505	.00908	.287
	(.092)	(.049)[18]	(.157)[14]					(.195)
(iv)	.432	.322	–	.059	–	.533	.00882	.370
	(.068)	(.053)[18]		(.025)[8]				(.157)
(v)	.502	.687	−.722	−.207	–	.603	.00814	.531
	(.146)	(.158)[24]	(.374)[18]	(.055)[10]				(.253)
(vi)	.136	.118	–	–	.0032	.432	.00974	.171
	(.030)	(.037)[12]			(.0034)[16]			(.179)
(vii)	.595	1.040	−2.179	−.187	.020	.758	.00635	−.085
	(.120)	(.172)[24]	(.499)[18]	(.099)[10]	(.012)[16]			(.250)

[a] Standard errors of coefficients (or sums of coefficients) are in parentheses, and distributed lag lengths are in brackets.

[b] $\bar{R}^2$ is the adjusted coefficient of determination; S_e is the standard error of estimate. The third column reports the coefficient and standard error attached to the lagged residuals of a regression in which the residuals from the ordinary least squares regression are regressed on the lagged residuals and the other explanatory variables of the model. A t-test on the coefficient reported in the third column can be used to test for the presence of first-order autocorrelation in the residuals. This test was proposed by James Durbin as an alternative to the use of his *h*-statistic when the latter is undefined which was the case in many of our equations.

combinations of lag lengths on orders and real wage rates, and selected the combination that minimized S_e. As the results indicate, this invariably produced a sum for the lag coefficients on real wage rates that has the correct sign and is highly significant. In addition, the point estimates on the adjustment speed, λ, rose substantially, and the overall statistical significance of both the adjustment speed and the response of normal orders improved.

A similar procedure was followed for normal real raw material prices. We added a distributed lag of real raw material prices to the basic model. The search process produced equation (iv) for each sector. This experiment was somewhat less successful in that only for Total Manufacturing was the sum of the lag weights on real raw material prices both significant and of the predicted sign. Nevertheless, in comparison with the results of equations (i) and (ii) the estimates of the speed of adjustment are again generally higher.

Finally, we ran regressions where distributed lags of both real wage rates and real raw material prices were added to the basic model. After a search process over various lag combinations, we settled on regression (v) for each sector. The results are very favorable. Each of the sum of the lag coefficients has the correct sign and is significant, often highly so. More important, the estimates of the speed of adjustment indicate very rapid adjustment. Indeed, the estimates of λ are insignificant from unity in both Total Manufacturing and Nondurables, indicating complete adjustment of actual to desired inventories within a quarter; in Durables, the point estimate is significantly higher than in the conventional literature, though it is still significantly below unity. Furthermore, observe that the lag lengths on the best-fitting lag combinations tend to be quite long indicating that normal variables respond slowly to changes in actual magnitudes.

In general, the results support the view that firms eliminate relatively quickly any gaps between desired and actual inventories, but that desired stocks are altered very slowly because the normal levels of orders and real factor-input prices that are the major determinants of desired stocks change slowly. This contrasts rather dramatically with the results in the conventional literature, typified by regression (i) in the table, where adjustment speeds tend to be very slow and where the lags in the adjustment of desired stocks tend to be fairly short. In our view the slow adjustment speeds that appear in the literature are due essentially to specification biases. The latter are caused by a combination of the exclusion of relevant explanatory variables, namely, normal factor-input prices, and the utilization of insufficiently long lags.

Next, we investigated the impact of real interest rates on inventory investment. We added a distributed lag of *ex post* real rates of interest as a measure of normal real interest rates, first to equation (ii) and subsequently to equation (v). After a search over various lag lengths on real interest rates, we arrived at equations (vi) and (vii) as the best-fitting equations. As one can see from the table, we were unable to uncover any strong effect of real interest rates on inventory investment. In every case, the sum of the lag coefficients on real interest rates is insignificant, and it frequently has the wrong sign. Furthermore, the overall fit of the equations, and the estimates of λ and the other parameters of the model are generally only marginally affected by the addition of real interest rates to the model.

To check these results, two other experiments were undertaken. First, we undertook a broader search than that involved in equations (vi) and (vii) in which the lag lengths on orders and factor-input prices as well as real rates were varied. Second, we decomposed the normal real rate into a normal nominal interest rate and a normal rate of inflation and fit separate distributed lags for each of the latter magnitudes. These experiments, however, did not change our conclusions regarding the effect of real interest rates or other variables on inventory investment, and hence the results of the experiments are not reported here.

In general, we find no strong, systematic influence of real interest rates on inventory investment. In this respect, we are no different from the conventional literature which has consistently failed to uncover an influence of interest rates on inventory investment at least in manufacturing industries.

III. Conclusions and Extensions

In this paper, we have used a modified flexible accelerator model to undertake an empirical analysis of finished goods inventory investment. Our major finding is that the speed with which actual inventories are adjusted to desired inventories is indeed very quick; we find in particular that the adjustment is essentially completed within a quarter in Total Manufacturing and Nondurables and is substantially completed within a quarter in Durables. This result, however, is obtained only in a properly specified model where desired stocks depend on the normal levels of both orders and real factor input prices and where these normal levels change slowly in response to changes in actual levels of the relevant variables. At the same time, we did not find any evidence that real interest rates have an important effect on inventory investment.

Two extensions of the model are needed. First, the model needs to be extended to consider the interaction of finished goods inventory investment with the accumulation of other stocks. These include unfilled orders, raw materials inventories, employment, etc. Secondly, the model needs to be estimated under expectation formation assumptions other than autoregressive expectations. We used autoregressive expectations in this paper because the inventory investment literature has generally used this assumption and we wanted to see whether a better explanation of inventory investment could be achieved with an improved specification of the determinants of inventory investment, given commonly used assumptions about expectation formation. Nevertheless, this is a restrictive assumption, and the model needs to be estimated under alternatives, for example, rational expectations, to check the robustness of our results.

REFERENCES

David Belsley, *Industry Production Behavior: The Order-Stock Distinction*, Amsterdam 1969.

J. Durbin, "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica*, May 1970, *38*, 410–21.

M. Feldstein, and A. Auerbach, "Inventory Behavior in Durable Goods Manufacturing: The Target Adjustment Model," *Brookings Papers*, Washington 1976, *2*, 351–96.

M. Lovell, "Manufacturers' Inventories, Sales Expectations, and the Acceleration Principle," *Econometrica*, July 1961, *29*, 293–314.

L. Maccini, "The Impact of Demand and Price Expectations on the Behavior of Prices," *Amer. Econ. Rev.*, Mar. 1978, *68*, 134–45.

J. C. R. Rowley, and P. K. Trivedi, *Econometrics of Investment*, New York 1975.

Merchant Wholesaler Inventory Investment and the Cost of Capital

By F. OWEN IRVINE, JR.*

We know from economic theory that the level of inventory a firm desires to hold should be inversely related to the per unit cost of holding the inventory. However, this theoretical relationship has generally eluded empirical verification. Following the repeated failures of nearly all earlier empirical studies to find any significant effect of interest rates on inventory investment, most recent studies have ignored the possible relationship between target inventory levels and inventory carrying costs.[1] Target inventory levels have been postulated to depend solely on expected sales.[2] Recently, however, Laura Rubin found aggregate (N.I.P.A.) inventory investment to be sensitive to fluctuations in financial inventory carrying costs. Also I have found in forthcoming papers that aggregate *U.S.* retail inventories and aggregate new automobile inventories depend in a statistically significant and economically important manner on a Hall-Jorgenson (H-J) service cost of capital measure.

This paper is an econometric study of merchant wholesaler (*MW*) inventories.[3] Merchant wholesales buy goods from producers and hold them in inventory for resale to their customers, who are either other producers or retail stores. Individual time-series equations explaining the monthly inventory levels held by this sector and by each of the subsectors listed in Table 1 (except farm products) are estimated. These inventories are found to be quite sensitive to fluctuations in financial inventory carrying costs.

I. Model Specification

Let $y(t)$ denote the actual amount of inventory on hand at the beginning of period t and $y^d(t)$ the optimum (target) level of inventories at the beginning of period t. The change in $y(t)$ over the period is assumed to consist of a planned and an unplanned part. I model the planned change as resulting from the firms partially adjusting their stocks toward their target levels. Changing the level of inventories involves per unit adjustment costs which generally increase with the amount of inventory change attempted per period. Hence firms spread their adjustment over several periods. The unplanned (or passive) inventory change we model as resulting from sales forecast errors.

$$(1)\quad y(t)-y(t-1)=\delta(y^d(t)-y(t-1)) \\ +\lambda(SF(t-1)-S(t-1))$$

where $SF(t-1)$ = sales forecast for period $t-1$, $S(t-1)$ = actual sales in period $t-1$, and δ = "speed-of-adjustment" coefficient. The greater the extent the firms can observe within the period that their sales expectations are incorrect and can take corrective actions (such as cancelling or increasing orders), the smaller will be the sales anticipation error coefficient, λ. Both λ and δ are expected to be between zero and one.

The unobservable target inventory $y^d(t)$ is postulated to be a function of expected future sales and expected inventory carrying costs:

$$(2)\quad y^d(t)=\alpha_0+\alpha_1\,{}_{t-1}\bar{S}(t)+\alpha_2\overline{CAPC}(t)$$

where ${}_{t-1}\bar{S}(t)$ = expected sales over the

*Department of economics, Wesleyan University. This paper was completed while I visited at the Federal Reserve Board as a Brookings Economic Policy Fellow. I want to thank Masahiro Kawai, Laura Rubin, and Che Tsao for helpful comments and Peter Ritter and Jeff Fuhrer for their excellent research assistance.

[1] See Michael Lovell (1964), fn. 4 of my forthcoming *American Economic Review* paper, and my 1981 paper for surveys of those studies which attempted to statistically verify this relationship.

[2] See the Michael Evans, J. L. Bridge, and J. C. R. Rowley and P. K. Trevidi inventory investment surveys.

[3] In the 1970's, these averaged about 18 percent of manufacturing and trade stocks.

TABLE 1—MERCHANT WHOLESALER SUBSECTOR DEFINITIONS

Subsector Name	Types of Goods Handled[a]	Jan. 1972 Stock[b]
Retail-Related Durables	Appliances & TV (39), Hardware (23), Home Furnishings (14), Jewelry (11)	3.06
Retail-Related Nondurables	Alcoholic Beverages (22), Drugs (17), Sporting Goods (17), Apparel (17)	4.14
Groceries	Groceries and Related Products (100)	2.64
Other Durables	Machinery & Equipment (34), Metals (14), Auto Equipment (13), Motor Vehicles (9), Electrical Goods (10), Lumber (9)	14.51
Other Nondurables	Paper (21), Piece Goods (13), Chemicals (11), Petroleum Products (9)	2.41
Farm Products	Farm Product Raw Materials (100)	2.11

Source: U.S. Census Bureau, *Monthly Wholesale Trade*, and unpublished data.

[a] Approximate percent of subsector.

[b] Shown in billion 1972 dollars.

inventory planning period conditional on information through period $t-1$, and $\overline{CAPC}(t)$ = expected per unit inventory carrying cost.

Substituting (2) into (1) and adding a stochastic error term yields

$$(3)\quad y(t)=(1-\delta)y(t-1)+\delta\alpha_0+\delta\alpha_1\,{}_{t-1}\bar{S}(t)+\delta\alpha_2\overline{CAPC}(t)+\lambda(SF(t-1)-S(t-1))+\varepsilon(t)$$

II. Data and Variable Specifications

Equation (3) was first estimated to explain monthly variations in the inventory level held by all merchant wholesalers excluding those carrying food (i.e., farm products and groceries). The dependent variable is the appropriately deflated beginning-of-the-month seasonally unadjusted inventory level held by these firms nationwide, $y_i(t)$. Since I felt that wholesalers selling finished goods to retail firms probably differ substantially from wholesalers selling intermediate goods to manufacturers, the subsectors reported in Table 1 were delineated. For these subsectors, inventory and sales series were constructed over 1967–75 from unpublished monthly data. (An appendix which reports all the data transformation details, subsector classification system, etc. is available from me upon request.)

To estimate equation (3), sales forecasts, ${}_{t-1}\bar{S}(t)$, are needed. Lacking observed data on expectations, I assumed the ${}_{t-1}\bar{S}(t)$ were generated by equation (4) which adjusts last year's sales in the same month by recent sales trends. Such a forecasting process is commonly utilized by trade sector firms and has been used successfully to model sales in several previous studies.[4]

$$(4)\quad SF_i(t)=DSALES_i(t-12)\times\left[\left(\frac{1}{3}\right)\left[\frac{DSALES_i(t-1)}{DSALES_i(t-13)}+\frac{DSALES_i(t-2)}{DSALES_i(t-14)}+\frac{DSALES_i(t-3)}{DSALES_i(t-15)}\right]\right]$$

where $SF_i(t)$ = expected deflated sales in month t for subsector i, and $DSALES_i(t)$ = actual deflated sales in month t for subsector i. In addition to $SF_i(t)$, equation (4) was utilized to generate sales forecasts for the next five months $[SF_i(t+1),\ldots,SF_i(t+5)]$ by substituting $[DSALES_i(t-11),\ldots,DSALES_i(t-7)]$ for the first term to the

[4] See my forthcoming papers for references.

right of the equals sign in (4). By initially including $SF_i(t),\ldots,SF_i(t+5)$ for ${}_{t-1}\bar{S}(t)$ in equation (3), I avoid any specification error in the absence of knowledge about the actual length of the inventory planning period. The correlations between monthly *changes* in deflated actual sales and $SF_i(t)$ over a 1968–74 sample are .70 for the total MW sector, .80 for Retail-Related Durables, .74 for Retail-Related Nondurables, .67 for Other Durables, and .53 for Other Nondurables.

Good-specific inventory carrying costs, which vary with the storage arrangements and depreciation rates of the goods inventoried, were not considered since they fluctuate very little over time. The capital costs associated with financing a firm's inventory or financial inventory carrying costs do vary considerably over time. They consist of two components. First, there is the opportunity cost of the funds invested in the inventory, which depends on the level of nominal interest rates. Second, there is the appreciation (or depreciation) in the price of the good while it is held in inventory. These "inventory profits" reduce the per unit financial inventory carrying costs. I used H-J service cost of capital variables, which capture both these components, as the primary measure of financial inventory carrying costs:

$$(5)\quad \overline{CAPC}_i(t)=\frac{P_i(t)}{PC(t)}\left[r(t)-{}^{e}_{t-1}DP_i(t)\right]$$

where $i=T$, RRD, $RRND$, G, OD, OND; $r(t)$ = short-term interest rate prevailing at beginning of period t; $PC(t)$ = Consumer Price Index (all items); $P_i(t)$ = producer price index of the ith subsector's goods; ${}^{e}_{t-1}DP_i(t)$ = average expected inflation rate of $P_i(t)$ over the holding period given information available through the end of period $t-1$. With $r(t)$ and the expected rate of inflation in annual percentage terms, $\overline{CAPC}_i(t)$ gives the number of real dollars per year it costs to hold a unit of inventory.

Since inventory is held a relatively short period of time, the beginning-of-the-period nominal interest rate is a good measure of the opportunity cost of funds invested in the marginal unit. The absence of observed price expectations presents a major problem, particularly since the actual inflation rate of the goods handled by merchant wholesalers fluctuated between a 2.5 and a 23.0 percent annual rate over the 1968–80 sample period. What matters to a wholesaler is the amount of inflation which occurs over the entire N-month holding period , not the month-to-month time path of inflation. Four types of models were utilized to generate forecasts of the average inflation rate over the next N months. Alternatively, the average expected rate of inflation was set equal to 1) the "naive expectation" that it would be the same as it was over the previous N months, 2) the average inflation rate forecast recursively by an $AR(2)$ model (the ARIMA model that provided the best fit to the second differences of the *log* of the price index), 3) to the average inflation forecast recursively by an equation explaining *log* first differences of the price level as a polynomial distributed lag (PDL) of the *log* changes in the previous nine periods,[5] and 4) the fitted value of the actual average rate of inflation which occurred in the subsequent N months. Under the "rational expectations hypothesis" the actual rate differs from the expected rate by a random error and hence can be substituted for the expected rate, provided the errors-in-variables problem this substitution creates is handled properly.

Since the value of N is unknown, N was set alternatively at 2, 6, and 12 months. Then after each alternative N-month inflation forecast (at an annual rate) was substituted into (5), the inventory equation (6) was estimated with each alternative capital cost measure, $CAPCDPN_i(t)$.

$$\begin{aligned}(6)\quad y_i(t)=&(1-\delta)y_i(t-1)+\delta\alpha_0\\&+\delta\sum_{j=0}^{5}\beta_jSF_i(t+j)\\&+\delta\alpha_2CAPCDPN_i(t)+\lambda(SF_i(t-1)\\&-DSALES_i(t-1))+\varepsilon(t)\end{aligned}$$

[5] The sum of the lag coefficients is .95 with a standard error of (.046). The mean lag is .85 months with the first three coefficients being .48, .28., and .13.

Notice that $r(t)$ and ${}_{t-1}^{e}DP_i(t)$ enter (5) with equal-sized coefficients. While this is perfectly logical under Jorgenson's model's assumptions, in a world where there is considerably more uncertainty about ${}_{t-1}^{e}DP_i(t)$ than there is about $r(t)$, risk-averting firms may weight ${}_{t-1}^{e}DP_i(t)$ less. For example, in a simple model of inventory speculation, Rowley and Trivedi show that the coefficient on ${}_{t-1}^{e}DP_i(t)$ in the target inventory equation is inversely related to the variance of the distribution of the expected inflation rates. Hence alternative versions of (6) were estimated with the prime interest rate and the expected inflation rate entering separately.

III. Estimation Results

Results for the total wholesale sector are given in Table 2 first for a 1968–74 sample, the period for which subsector data was available, and then for a sample extended through 1980. The sales forecast for month $t+5$ was statistically insignificant and was dropped, which suggests that the inventory planning period is about five months. The F-tests reject the hypotheses that $\beta_3=\beta_4=\beta_5$ and strongly reject the hypothesis that $\beta_0=\beta_1=\beta_2=\beta_3=\beta_4=0$. The point-of-means elasticities of the target inventory level with respect to ${}_{t-1}\overline{S}(t)$, given at the bottom of Table 2, are generally within the expected range.

For the 1968–74 sample, a H-J capital cost based on naive inflation expectations produced the best fit (eq. (7)). The one utilizing the *PDL* 6-month-ahead forecast of inflation had the next best fit (equation (8)). The elasticities of y^d with respect to these are within the $-.05$ to $-.07$ range found previously for aggregate *U.S.* retail inventories (see my forthcoming paper). However, as equations (9) and (10) show, splitting *CAPCDPN* into the prime rate and the expected inflation measures: 1) significantly improves the fit; 2) suggests that naive expectations are the better model over this sample; and 3) shows that wholesalers tend to place much more weight on the nominal interest rate than on inflation expectations. Assuming rational expectations, the actual inflation rate is treated as endogenous in equation (11). This causes the fit to worsen.

Over the 1968–80 sample, in an equation not reported, the estimated coefficient on a H-J capital cost measure assuming naive expectations is near zero and has a .05 t-statistic. Of the other H-J measures, the one based on the 6-month-ahead *PDL* forecast again produced the best fit, but has a statistically insignificant coefficient (equation 12). From equations (13) and (14) it is clear that the reason for this deterioration is that the inflation forecasts no longer enter significantly. The actual future inflation rate (assuming rational expectations) also had a near zero statistically insignificant coefficient. This failure to find a statistically significant effect of expected inflation on inventory investment over the extended sample probably should be interpreted as a failure of the models of expected inflation discussed earlier to characterize actual inflation expectations, rather than as evidence that wholesalers ignored the roller coaster inflation of the late 1970's. On the other hand, the coefficients on the prime interest rate are statistically very significant and imply very large elasticities.

Turning to the subsector equations reported in Table 3, we find that a H-J capital cost measure based on naive expectations has a negative and significant coefficient in the three retail-related subsectors.[6] It also enters with a negative and reasonably significant coefficient in the large Other Durables subsector. Groceries, which turn over quickly due to high good-specific carrying costs, have the smallest capital cost elasticity as expected.

The coefficients show that retail-related inventories are sensitive to sales forecasts only over the next three months, in contrast to the dependence of nonretail-related inventories on the forecasts of sales two quarters ahead. This suggests different length inventory planning periods. Also, the estimated speeds of adjustment are gener-

[6] The H-J capital cost measures assuming rational expectations also entered with appropriate signs for the Retail-Related Durables and Nondurables sectors, but produced worse fits than the equations reported in Table 3.

TABLE 2—INVENTORY EQUATIONS FOR TOTAL MERCHANT WHOLESALER SECTOR EXCLUDING FOOD[a]

	June 1968 through May 1974					June 1968 through Jan. 1980		
Coefficient of:	(7)	(8)	(9)	(10)	(11)[b]	(12)	(13)	(14)
Inventory Level $(t-1)$	.921	.924	.898	.900	.899	.971	.971	.972
	(.028)	(.039)	(.031)	(.032)	(.039)	(.010)	(.010)	(.010)
Sales Expected in	.012	.013	.026	.027	.017	.040	.049	0.48
Month t	(.024)	(.025)	(.025)	(.026)	(.028)	(.020)	(.020)	(.020)
Month $t+1$	.032	.033	.036	.038	.041	−.020	−.019	−.019
	(.026)	(.026)	(.026)	(.026)	(.026)	(.020)	(.020)	(.020)
Month $t+2$	−.020	−.023	−.018	−.020	−.021	−.007	−.005	−.005
	(.025)	(.025)	(.025)	(.025)	(.026)	(.019)	(.019)	(.019)
Month $t+3$	−.001	−.001	.005	.006	.008	−.007	−.004	−.004
	(.025)	(.025)	(.025)	(.026)	(.026)	(.019)	(.019)	(.018)
Month $t+4$	.057	.057	.064	.064	.044	.054	.058	.058
	(.022)	(.022)	(.022)	(.022)	(.026)	(.018)	(.017)	(.017)
Sum of Above Coefficients on Sales	.079	.082	.113	.115	.090	.059	.079	.078
Previous Month's Sales	.007	.002	.014	.011	.020	−.005	.006	.008
Forecast Error	(.034)	(.034)	(.034)	(.034)	(.036)	(.027)	(.028)	(.027)
Service Cost of Capital with								
Expected Inflation Measured by:								
Actual Inflation Previous Year,	−.039							
CAPCDP12N	(.017)							
Six-Month Ahead Forecast		−.030				−.020		
from *PDL*, *CAPCDP6PDL*		(.016)				(.019)		
Prime Bank Interest Rate			−.059	−.051	−.089		−.048	−.047
			(.020)	(.020)	(.045)		(.020)	(.020)
Expected Inflation as Measured by:								
Actual Inflation Previous Year,			.028				.003	
FDP12N			(.015)				(.009)	
Six-Month Ahead Forecast				.018				.003
from *PDL*, *FDP6PDL*				(.015)				(.017)
Actual Inflation Over Next Six					.023			
Months, *FDP6A* (endogenous)					(.018)			
Constant Term	.828	.732	.956	.848	1.608	−.153	−.324	−.328
	(.355)	(.368)	(.349)	(.355)	(.862)	(.246)	(.218)	(.241)
RHO	−.089	−.060	−.100	−.073	.090	.202	.162	.162
R^2	.9951	.9950	.9953	.9952	.9946	.9979	.9980	.9980
Standard Error	.2063	.2090	.2033	.2061	.2195	.2834	.2777	.2778
R^2 in Terms of Changes	.2251	.2041	.2595	.2391	.1373	.1798	.2184	.2179
Durbin-Watson Statistic	1.96	1.97	1.97	1.98	2.01	2.01	2.01	2.01
Estimated Speed of Adjustment, δ	.079	.076	.102	.100	.101	.029	.029	.028
Point-of-Means Elasticity of y^d with respect to:								
Sum of Future Expected Sales	.738	.796	.817	.848	.657	1.343	1.861	1.903
Service Cost-of-Capital Measure	−.060	−.070				−.086		
Prime Bank Interest Rate			−.147	−.130	−.224		−.403	−.405
Expected Inflation Measure			.033	.044	.045		.019	.012

[a] Equations estimated by instrumental variables with the lagged dependent variable treated as endogenous in order to obtain consistent coefficient estimates. Fair's method was utilized to correct for serial correlation. *RHO* is the final value of the autocorrelation parameter found after several iterations. Asymptotic standard errors are reported in the parentheses. The estimated coefficients of the target inventory equation (2), which were derived by dividing coefficients above by the estimated δ, were used to calculate the point-of-means elasticities.

[b] The six-month-ahead *PDL* inflation forecast was added as an excluded instrument to this equation.

TABLE 3—INVENTORY EQUATIONS FOR MERCHANT WHOLESALER SUBSECTORS

Coefficient of:	Retail-Related Durables 1968:6–75:3 (15)	Retail-Related Nondurables 1968:6–74:5 (16)	Groceries 1968:6–75:3 (17)	Other Nondurables 1968:6–74:5 (18)	Other Durables 1968:6–74:5 (19)
Inventory Level $(t-1)$	.208	.646	.489	.606	.983
	(.152)	(.070)	(.126)	(.125)	(.069)
Sales Expected in					
Month t	.292		.045		.014
	(.069)		(.043)		(.035)
Month $t+1$	.179	.131	.062	.105	−.045
	(.064)	(.055)	(.041)	(.051)	(.038)
Month $t+2$	.165	.315	.089		−.001
	(.057)	(.041)	(.039)		(.035)
Sum of Months $t+3$, $t+4$, and $t+5$				.064	.134
				(.027)	(.121)
Sum of Above Coefficients on Sales	.636	.446	.195	.169	.102
Previous Month's Sales Forecast Error	−.191	−.027	.014		.036
	(.097)	(.065)	(.048)		(.048)
Capital Cost with Expected Inflation Measured by Previous Year's Actual Rate	−.021	−.040	−.0078	.010	−.018
	(.006)	(.014)	(.0025)	(.003)	(.015)
Constant Term	1.453	1.294	4.434	2.359	2.632
	(.328)	(1.677)	(2.848)	(1.856)	(5.306)
RHO	.597	.253	.253	.391	−.028
R^2	.712	.949	.725	.846	.990
Standard Error	.0702	.0802	.0787	.0693	.1339
Durbin-Watson Statistic	1.22	1.63	1.58	1.22	1.96
Estimated Speed of Adjustment, δ	.792	.354	.511	.394	.017
Point-of-Means Elasticities of y^d with respect to:					
Sum of Expected Future Sales	.47[a]	1.03	.678	.769	.370
Capital Cost Measure	−.03[a]	−.103	−.0063	.020	−.192

[a] See Table 2.

ally faster for the retail-related subsectors. This variation in speeds of adjustment, in the length of the inventory planning horizons and in the sensitivity to capital costs, supports the a priori expectation that there would be benefits to disaggregation.

IV. Implications and Further Work

The most important finding of this study is that *MW* inventories depend in both a statistically significant and an economically important manner on the level of financial inventory carrying costs. The negative coefficient on the prime interest rate is statistically very significant in both sample periods. Its elasticity is also at least three times larger than the elasticity with respect to the expected inflation measures, whose coefficients are also less statistically significant. The estimates imply that a one-point rise in the prime rate, *ceteris paribus*, will decrease target *MW* inventories (excluding food) by from $0.6 to 1.7 billion 1972 dollars. These inventories averaged $33 billion 1972 dollars over 1968–80. Hence these estimates imply that the cyclical swings in short-term interest rates of five points or more, which have been common over the 1970's, caused target wholesale stocks to change from about 10 to 25 percent. Thus, changes in monetary policy which alter short-term interest rates do have a significant impact on wholesale inventory investment. This channel by which monetary policy influences the real economy is potentially an important one for stabilization policy since the lags between changes in short-term interest rates and inventory investment are short. Also, as Alan Blinder and Stanley Fischer point out, it is a channel through which even anticipated money supply changes can influence the real economy in the presence of rational expectations.

There are several obvious avenues for further research suggested by these results. The dependence of inventory targets on sales and inflation expectations and the ability of inventory levels to respond quickly to changes in these expectations suggests that increased understanding of the process by which expectations are formed is crucial. Second, the results suggest that there are benefits to disaggregation. Finally, the statistical relationship between manufacturing inventory investment and capital costs needs further investigation in light of findings that retail and merchant wholesaler inventories depend significantly on capital cost fluctuations.

REFERENCES

A. S. Blinder and S. Fischer, "Inventories, Rational Expectations, and the Business Cycle," *J. Monet. Econ.*, forthcoming.

J. L. Bridge, *Applied Econometrics*, Amsterdam 1971.

M. K. Evans, "Inventory Investment," in *Macroeconomic Activity*, New York 1969.

F. O. Irvine, "Retail Inventory Investment and the Cost of Capital," *Amer. Econ. Rev.*, forthcoming.

______, "A Study of Automobile Inventory Investment," NI-WPP work. paper 6, Federal Reserve Board, Dec. 1971.

______, "The Dependence of Aggregate Inventory Investment on Inventory Carrying Costs, A Critique of Recent Research," *Proc. First International Symposium on Inventories*, Akademiai Kiado, Budapest 1981.

Mi. C. Lovell, "Determinants of Inventory Behavior," in Edward F. Denison and Lawrence R. Klein, eds., *Models of Income Determination*, Nat. Bur. of Econ. Res. *Stud. in Income and Wealth*, Vol. 28, New York 1964.

J. C. R. Rowley and P. K. Trivedi, *Econometrics of Investment*, New York 1975.

L. S. Rubin, "Aggregate Inventory Behavior: Its Response to Uncertainty and Interest Rates," *J. Post Keynesian Econ.*, Winter 1979–80, *2*, 201–11.

U.S. Census Bureau, *Monthly Wholesale Trade*, various years, 1967–80.

Capital-Labor Conflict and the Productivity Slowdown

By DAVID M. GORDON*

> [Management] got all the technological improvements... But one thing went wrong... We've been telling them since we've been here: We have a say in how hard we're going to work. They didn't believe us.
>
> *U.S. autoworker, ca. 1973*
> [quoted in Studs Terkel, p. 263]

The slowdown in productivity growth during the 1970's continues to puzzle neoclassical economists. As Edward Denison concludes, "what happened is, to be blunt, a mystery"(p. 4).

This mystification should not seem particularly surprising, since neoclassical economists pay so little attention to the operations of the process of production. In contrast, Marxist economists have recently placed high priority on analyses of production relations in capitalist economies (see Herbert Gintis; and Michael Reich and James Devine). Although not yet applied to studies of aggregate labor productivity, this work should nonetheless provide fruitful guidelines for Marxian investigations of the productivity slowdown.

Among other suggestions, the recent literature emphasizes the importance of both *external* and *internal* mechanisms of labor control. Where would these clues guide our initial explorations of the productivity puzzle?

The external effect seems clearly to point in the *wrong* direction. Labor markets have become relatively looser during the 1970's. This heightened labor market competition ought to have pushed workers harder and, other things equal, to have *increased* labor productivity.

In contrast, the clues about internal control mechanisms seem more promising. Many corporations have been complaining about worker performance, while both absenteeism and worker dissatisfaction have been rising (see Graham Staines).

Pursuing these clues, I develop and empirically test in this paper a formal Marxian model of aggregate labor productivity which pays special attention to both the emergence and erosion in the postwar *U.S.* economy of a vast internal corporate apparatus of "bureaucratic control." (See Richard C. Edwards for definition and elaboration.) The analysis is very provisional, but the results reported here nonetheless provide strong support for a single central conclusion: The declining effectiveness of the postwar system of bureaucratic control appears to explain *almost all* of the recent slowdown in productivity growth. Viewed from the Marxian perspective, the productivity slowdown seems no more mysterious than any other contradiction of capitalist economies.

I. A Marxian Model of Aggregate Labor Productivity in the United States

The Marxian analysis of aggregate labor productivity in capitalist economies begins with a fundamental theoretical distinction between labor power and labor activity. (See, for instance, Gintis.) This distinction suggests that aggregate output must be analyzed as a function *both* of factor potential *and* of the degree to which capitalists succeed in extracting as much surplus labor as they can from the labor power they have purchased in the labor market. Marxists also argue

*Graduate Faculty, New School for Social Research. Thanks are due Martin N. Baily, Samuel Bowles, Robert J. Gordon, Michele Naples, Michael Reich, Howard Sherman, Thomas E. Weisskopf, and colleagues at the New School for helpful comments on an earlier draft. I owe a greater debt to workers whom I have met in labor education projects; their angry and insightful comments about their bosses and supervisors initially suggested this entire project.

theoretically that only production workers —and not other nonproduction workers—directly contribute to the actual use-value of goods and services produced in capitalist economies.

These two introductory guidelines point directly toward a formal Marxian framework for the analysis of aggregate labor productivity. We can postulate in general that

$$(1) \qquad y^p = p^p \cdot i^p$$

where the superscripts remind us that the analysis is framed exclusively for production workers; y^p is total aggregate output per hour of production-worker labor power purchased; p^p is average (aggregate) potential output per hour of purchased production-worker labor power; and i^p is the average (aggregate) intensity of concrete labor effort in production, with a potential range of $0 \leqslant i^p \leqslant 1$.

The Marxian analysis of the determinants of p^p is parallel to neoclassical analyses: p^p is likely to increase if production workers are able to combine their labor with relatively more *capital goods*, *raw materials* and *embodied labor skills*.

The general determination of i^p requires a historical perspective. My paper with Edwards and Reich argues that each successive stage of capitalist development in the United States has featured a new (and supplementary) mechanism affecting the average intensity of labor effort:

Since the middle of the nineteenth century in the United States, the generalization of the wage-labor market—through a process which Marxists call "proletarianization"—has created an external mechanism of labor discipline: the reserve pool of wage-labor. The greater the threat of dismissal and unemployment, the more compliant (and productive!) currently employed production workers are likely to become.

In our paper we also postulate that, since the turn of the century, corporations have sought to increase their external and internal leverage over workers through strategies of "divide-and-conquer": The greater the divisions among wage-laborers, the weaker will be production workers' resistance on the job.

Since the Great Depression and World War II, systems of "bureaucratic control" have emerged as a principal addition mechanism through which large corporations monitor and seek to manipulate worker effort. (The ratio of nonproduction to production labor increased by 71 percent from 1947 to 1968.) A *quid pro quo* grounded the "labor peace" of this period: Workers and their unions would tolerate an intensification of supervisory and administrative control over production *in exchange for* a steadily increasing real wage and improving working conditions.

Combining these three hypotheses, we can postulate a simple functional relationship for the postwar *U.S.* economy: i^p will vary directly with r, the relative size of the reserve pool of production workers; d, the intensity of divisions in the wage-labor force; and b, the relative effectiveness of bureaucratic control.

One final general hypothesis flows from the Marxian emphasis on capital-labor conflict. Potential output is, after all, only *potential* and not *actual*. It is therefore likely, particulary in the short and medium run, that variations in the elements of i^p will explain a substantially greater portion of the variance over time in y^p than will variations in the elements of p^p.

II. A Comparative Empirical Analysis of Labor Productivity Growth

I have adopted some relatively unrestrictive specifications of the general model outlined above in order to facilitate comparison with traditional neoclassical models and to moderate potential econometric problems.

We can assume that p^p will increase with the increase in 1) k^p and x^p, the value of capital goods and energy (respectively) available per hour of purchased production-worker labor power; and 2) v^p, an index of labor skills, based on educational attainment, embodied in production-worker labor power. (I am ignoring, for the moment, the weakness of schooling as a proxy.)

I postulate that the reserve labor effect r operates with a lag and is better measured by the inverse of the quit rate than by the unemployment rate (since the former more directly traces the behavior of *currently* employed workers). The greater the decline in the quit rate q in the previous period (and, therefore, the looser the labor market), *ceteris paribus*, the greater will be the increase in labor productivity. (See related recent work by Gerry Oster, and by David Stern and Daniel Friedman.)

In order to measure most inclusively the effects of labor divisions, one would need an aggregate index of the (weighted) inequalities among all competing groups in the wage-labor force. Since I have not had time to construct such a composite index for this first stage of analysis, I have relied on an incomplete substitute: Denison's index of the changing age-sex composition of the labor force. Contrary to reigning neoclassical expectations, I hypothesize that labor productivity will *increase* with rising relative labor force participation of women and teenage workers because of heighted labor-force divisions and the relatively weaker bargaining strength, at least in this period, of women and teen-age workers.

Adequate specification of b is obviously complicated. At this initial stage of investigation, I postulate that labor productivity will increase directly with a rising intensity of supervision and administration *but* that the relative effectiveness of the postwar labor-peace bargain will condition this productivity-enhancing effect of the apparatus of bureaucratic control. This suggests the following formulation:

$$(2) \qquad Db = \alpha_1 D[w^*/n^*] + \alpha_2 Dn$$

where D indicates the exponential rate of change of the respective variables; n is the ratio of nonproduction to production employee hours (adjusted for "hoarding" over the cycle); w^* is the deviation of total (real) production-worker compensation around its trend; and n^* is the deviation of n around its trend. This specification suggests that the productivity effects of a rising n are likely to be *augmented* when real wages have increased sufficiently to induce production workers' cooperation with the growing administrative machine; and , conversely, that the productivity effects are likely to be *diminished* if current increases in the intensity of supervision are not adequately compensated.

We can translate the preceding discussion into a fully specified equation:

$$\begin{aligned}(3) \quad Dy^p &= g + a_1 Dv^p + a_2 Dk^p \\ &+ a_3 Dx^p - a_4 Dq^p_{t-1} + a_5 Dd \\ &+ a_6 D[w^*/n^*] + a_7 Dn + u_1\end{aligned}$$

where g is a constant rate of growth in y^p.

For purposes of comparison, we can also estimate the neoclassical equivalent:

$$\begin{aligned}(4) \quad Dy^T &= g + b_1 Dv^T + b_2 Dk^T + b_3 Dx^T \\ &- b_4 D[Y/Y^*] + u_2\end{aligned}$$

where the superscript T indicates that a variable is calculated in terms of total employee hours (not production-worker hours); v^T is now measured by Denison's composite index of "labor-quality"; and $[Y/Y^*]$ is the ratio of actual to potential output. We expect b_4 to have a negative sign following recent neoclassical hypotheses about the procyclical behavior of productivity: employers will "overadjust" to a rising $[Y/Y^*]$, eroding labor productivity.

Data were available at the time of estimation for all of the variables in equations (3) and (4) on an annual basis for the *U.S.* economy from 1947 to 1978. (Lags and first differences push the first observation to 1949.) I have estimated the equations initially for the nonfarm private sector, although future work will obviously require more disaggregated analysis. Table 1 presents the basic econometric results.

Column (1) reports the results for the neoclassical equation (4). It points toward familiar conclusions and puzzles. The capital, energy, and output gap variables confirm conventional hypotheses. (The labor quality variable performs perversely, a puzzle for human capital theory.) Consistent with results from Denison and others, the residuals of the estimated equation rise

TABLE 1—ALTERNATIVE EMPIRICAL MODELS OF AGGREGATE LABOR PRODUCTIVITY[a]

	Independent Variable		Neoclassical Equation (4) (1)	Marxian Equation (3) (2)
	g	Constant rate of growth	[d]	[d]
	Dv	Labor "quality" growth	-1.997[b] (1.882)	[d]
	Dk	Capital-labor growth	0.980[c] (4.068)	[d]
	Dx	Energy-labor growth	0.540* (2.286)	[d]
	$D[Y/Y^*]$	Output gap adjustment	-1.002[c] (3.060)	[e]
	Dq_{t-1}	Reserve labor effect	[e]	-0.428[c] (2.572)
	Dd	Labor divisions effect	[e]	[d]
Db	$D[w^*/n^*]$	Effectiveness of control	[e]	1.192[c] (3.541)
	Dn	Intensity of control	[e]	1.110[c] (2.725)
		R^2	0.560	0.749
		F-Statistic for equation	7.950[c]	7.227[c]
		Durbin-Watson statistic	2.165	1.672

Sources: See Data Appendix.

Note: T-statistics (absolute values) are in parentheses below coefficients. *Dependent Variable*: y^T or y^p (average annual change in output per hour); *Units of Observation*: Annual data for *U.S.* nonfarm private sector.

[a] All variables measured as exponential rates of change. All variables in column (1) defined in terms of total employee hours; all variables in column (2) defined in terms of production-worker hours.

[b] significant at 5 percent;

[c] significant at 1 percent (one-tail tests).

[d] coefficient less than its standard error.

[e] variable not estimated in equation.

steadily from 1951 to 1973 ($r_{e_t . time}=0.63$) before behaving unpredictably and rising substantially in average absolute value during 1974–78.

The Marxian analysis of the postwar period suggests that equation (3) should be applied to the period beginning in 1954, since we argue that the "labor peace" was not fully in place until the end of the Korean War. Column (2) presents the regression results for equation (3) for 1954–78.

The results provide striking initial support for the Marxian model. The coefficients on both the reserve labor and the bureaucratic control variables are all significant at 1 percent with the expected signs. Since the labor divisions variable incompletely measures the hypothesis, its insignificance is not surprising (although it did have the expected sign). The capital, energy, and labor quality variables are all insignificant, hinting at under specification bias in conventional neoclassical models. The elements of i^p account for 90 percent of total explained variance in the estimated equation, confirming our prior hypotheses about the relative importance of the elements of i^p and p^p.

Further, the results also suggest that equation (3) is able *both* to capture the basic structure of the postwar economy *and* to account for its "anomalous" behavior after 1973: On the one hand, the results are comparable if estimated for the period from 1954 to 1973; while, on the other hand, in the estimated equation in column (2), the absolute value of the residuals for 1954–73.

(There is also no apparent time trend in the residuals for 1954–73 $[r_{e_t \cdot time} = 0.07]$, raising questions about Denison's well-known "miscellaneous advances in knowledge and *n.e.*c.")[1]

We can perform a final econometric experiment to compare the adequacy of Marxian and neoclassical explanations of the productivity slowdown. Equations (3) and (4) were estimated for the period 1949–73 (a somewhat disadvantageous terrain, as suggested above, for the Marxian model). Those results were then used to forecast predicted rates of productivity change for 1974–78 and to compare predicted with actual rates of productivity growth. The neoclassical equation (4) forecasts as badly as the sirens of mystification would lead us to expect: The regression coefficient of actual on predicted productivity change is only 0.26 and the R^2 of the simple regression equation is 0.12. The Marxian equation (3) forecasts with notably greater accuracy: The regression coefficient of actual on predicted productivity change is 0.92, very close to a "perfect" forecast of 1.00; and the R^2 of the simple regression is 0.76.

III. Conclusions and Implications

What does the Marxian model tell us about the *sources* of declining productivity growth after 1973? Suppose that we assume that $D[w^*/n^*]$, the index of the effectiveness of bureaucratic control, had continued to grow from 1974 to 1978 at its trend value for the 1954-73 period. The results reported in column (2) would predict that the average annual growth in labor productivity would have been 1.4 percentage points higher in 1974–78 than its actual values. *This effect accounts for* 87.5 *percent of the measured retardation in labor productivity growth between* 1954–73 *and* 1974–78.

The underlying data reveal the sources of this erosion in the effectiveness of bureaucratic control. After the early 1970's it appears that corporations began to wage what UAW president Douglas Fraser has recently called a "one-sided class war," effectively abrogating the "labor peace" of the earlier period. One strategy of attack involved a dramatic resumption in the growth of the internal administrative apparatus —after an ebb from 1969–73—despite the profit squeeze and the escalating costs of administrative personnel. Another strategy apparently involved much tougher bargaining over money wages. Compounded by the effects of inflation, the net result was a dramatic decline in $D[w^*/n^*]$ resulting from *both* falling w^* *and* rising n^*.

This analysis does not "blame" either corporations or workers for the slowdown in productivity. It suggests that a *system* of bureaucratic control worked for a while and has since become counterproductive. Assuming that the *U.S.* economy is substantially restructured during the 1980's, we face a critical choice: Will internal corporate systems of control become even more centralized? Or shall we finally begin to move toward more participatory and democratic systems of worker coordination and self-managment?

Data Appendix

y^T, y^P: Output in 1972 dollars from *Survey of Current Business* adjusted for government purchases. Hours from *Business Conditions Digest*, with production-worker hours based on adjustment for nonproduction worker employment from *Employment and Training Report of the President.*

[1] Initial responses to these results prompt three additional comments:(a) Clerical workers are almost always counted as production workers in the variable n; nonproduction workers include, almost exclusively, supervisors and managers. (b) The (lagged) wage rate is obvious endogenous to this analysis; further exploration requires simultaneous (or at least instrumental) estimation. For the moment, it is useful simply to emphasize that the variable $[w^*/n^*]$ reflects substantial movement in both its numerator and its denominator; it is not perfectly correlated with the real wage. (c) Why is $D[Y/Y^*]$ or an alternative measure of capacity utilization (Dz) not included in equation (3)? The Marxian model controls both for the capital-labor ratio and for labor hoarding of nonproduction workers over the cycle; these are likely to be the principal reasons that labor productivity would vary with the output gap or Dz. In any case, I estimated seperate equations with $D[Y/Y^*]$ and Dz (respectively) added to equation (3); neither additional variable was statistically significant.

k, x: Both variables in constant dollars: capital from Denison; energy from *Historical Statistics of the United States* and the *Statistical Abstract*.

v, d: Based on series reported in Denison.

$[Y/Y^*]$: Based on series constructed by Jeffrey Perloff and Michael Wachter, extrapolated and generously provided by Robert Gordon.

q, w: Quit rate for manufacturing used as proxy for nonfarm business sector; from *Employment and Earnings, United States, 1909–1978*. Money wages in constant dollars, adjusted for overtime, interindustry shifts, and compensation, from *Economic Report of the President*.

n: Nonproduction worker hours based on series in *Employment and Training Report of the President*, translated into hours, adjusted for capacity utilization to reflect changes net of labor hoarding through cycle.

REFERENCES

Edward F. Denison, *Accounting for Slower Economic Growth*, Washington 1979.

Richard C. Edwards, *Contested Terrain*, New York 1979.

D. M. Gordon, R. C. Edwards, and M. Reich, "The Historical Development of Labor Segmentation in the United States," in *The Segmentation of Labor in the United States*, Cambridge 1981, forthcoming.

H. Gintis, "The Nature of Labor Exchange and the Theory of Capitalist Production," *Rev. Radical Polit. Econ.*, Summer 1976, *8*, 36–54.

G. Oster, "Labor Relations and Demand Relations: A Case Study of the 'Unemployment Effect'," *Cambridge J. Econ.*, forthcoming.

M. Reich and J. Devine, "The Microeconomics of Conflict and Hierarchy in Capitalist Production," *Rev. Radical Polit. Econ.*, Winter 1981, *13*.

G. L. Staines, "Is Worker Dissatisfaction Rising?," *Challenge*, May/June 1979, *22*, 38–45.

D. Stern and D. Friedman, "Short-Run Behavior of Labor Productivity: Tests of the Motivation Hypothesis," *J. Behav. Econ.*, forthcoming.

Studs Terkel, *Working*, New York 1974.

Industrial Conflict and its Implications for Productivity Growth

By MICHELE I. NAPLES*

In the last decade, economists have come to pay increasing attention to the behavior of productivity growth. Most have been concerned with the apparent secular drop in productivity growth in the 1970's. In addition, in a recent article, Robert Gordon pointed out a recurring anomaly — the slowdown in the rate of growth of productivity at the end of business expansions.

Although neoclassical economists have developed and explored a variety of explanations for the behavior of productivity growth in the postwar period, no one has investigated the extent to which the quality of industrial relations may have affected the growth of output per production worker-hour.

Marxian economists, on the other hand, take relations between labor and capital as the starting point of their economic analyses. Four years before Gordon's article was published, Raford Boddy and James Crotty predicted just such an end-of-expansion slowdown in productivity growth due to greater worker strength on the job when unemployment is at a minimum.

This suggests that greater attention to the quality of industrial relations, or put differently, to the extent and character of industrial conflict, could enable us to develop a better understanding of the behavior of productivity growth over the last three decades.

Although there has been some research on the question of industrial conflict, it is fairly limited. While a number of Marxian economists have inferred that conflict or labor militance is the causal link between unemployment and productivity growth, or unemployment and wage growth, no one has explicitly investigated this relationship empirically. A number of neoclassical analysts have examined the behavior of the volume and incidence of strikes (percent of days lost, percent of labor force striking), and of quit rates (see, for example, Orley Ashenfelter and George Johnson; John Pencavel). But I will argue that there are a number of dimensions of industrial conflict which cannot be captured by these measures alone.

This paper will outline hypotheses about the character of industrial conflict and will draw on these to develop measures of conflict which can be used to explore its relationship to productivity growth. It will also suggest hypotheses about the behavior of industrial conflict and its economic effects, although an empirical investigation of these two sets of hypotheses has not yet been completed.

I. Theoretical Propositions

I propose three sets of hypotheses regarding industrial conflict: 1) industrial conflict has several different dimensions which can be distinguished theoretically and empirically; 2) these different dimensions have different patterns of behavior over time; 3) these different dimensions impact on productivity growth to different extents.

First, industrial conflict can be divided into two categories: actions *individual workers* take which are an indication of conflict and actions taken by *groups of workers*. When a firm faces a high quit rate or absenteeism rate, it is forced to recognize employee dissatisfaction with working conditions as well as wages. But whereas such individual actions would naturally increase under tight labor markets when the expected cost of losing a job is least, the presence or absence of collective actions

*Assistant professor, economics and women's studies, Douglass College, Rutgers University. I would like to thank all who offered comments on earlier drafts, especially Sam Bowles, Jim Crotty, Gerry Duguay, David Gordon, and Kon Jo Wu.

would depend on the shape of the union movement as well as on labor market conditions.

In order to investigate the willingness of labor unions to be militant, it is necessary to examine the pattern of industrial relations established around the time of the AFL-CIO merger. That pattern can be characterized as a truce between unions and companies because it included gains for and concessions by both. Most unions curtailed their use of the strike, which was especially novel for the CIO unions. They learned to live with the return of the injunction (outlawed since 1932) embodied in the Taft-Hartley Act. At the same time unions did make gains: they achieved some financial security (for example, the deduction of union dues from paychecks), were recognized legally and accepted by corporations, and were able to obtain certain contract provisions (for example, the seniority principle and grievance mechanism) that helped protect workers from arbitrary treatment on the job. Of all of the aspects of this truce, two are most critical for an investigation of industrial conflict and its effects on productivity: the first has to do with the form of labor militance, the second with its object.

First, the new orientation towards industrial peace meant that unions came to rely more on expert negotiation to mediate between workers and employers, and reverted to the use of the strike only as a very last resort. If the possibility arose that a strike might erupt at the end of a contract, unions were required by law to give the company sixty days notice of that possibility. But during the term of a contract there was no disruption of production—for the most part grievance procedures were adopted to handle complaints about changes in the work process. Furthermore, the union was responsible for ensuring that the rank and file adhered to the contract, which implied disciplining recalcitrants when necessary.

The second important facet of the truce was that unions came to concentrate on income issues, and allowed corporations to exercise their freedom of enterprise with few restraints. In many cases, of course, union members benefited directly from technical changes and reorganizations of the workplace by having their wages tied to productivity growth. But any conflict over plant administration per se was relegated to slow grievance procedures or local union bargaining.

I hypothesize, in keeping with the thinking of a variety of other analysts of this period (Stanley Aronowitz, Samuel Bowles and Herbert Gintis, Murray Gart, and William Simkin), that this basic pattern for industrial relations, while certainly predominant in the 1950's, broke down in the long expansion of the 1960's. The contradictions embodied in such a truce made it less and less tenable: as union members became concerned that union officials were not satisfactorily representing their interests, they began to take matters into their own hands. As the number of shop stewards per worker fell and the time it took to process grievances increased, that mechanism for dealing with workplace issues became delegitimized. As the injury-frequency rate in manufacturing rose by more than 25 percent in the late 1960's into the 1970's (see *Statistical Abstract of the United States*), unionized workers began insisting that working conditions be included among contract demands.

As can be seen from Table 1, the proportion of strikes over working conditions increased secularly in this period, as did the proportion of strikes during the term of the contract (a proxy for wildcat strikes), and the proportion of wildcats which were concerned with working conditions. Also the proportion of tentative negotiated settlements rejected by rank-and-file vote jumped by more than half from 1964 to 1967, and remained high during economic expansions through the 1970's (see Federal Mediation and Conciliation Service).

I would thus expect to discern two very different patterns of aggressive labor militance over this period—one within the boundaries of the truce, and the other outside those boundaries. In addition to these two types of labor militance, there are times when collective actions are more an expression of labor's weakness than labor strength. Often, during slow times, businesses push

TABLE 1—THE BEHAVIOR OF THREE ELEMENTS OF MILITANCE, MINING AND MANUFACTURING, 1953–77

Indicator	1953–60	1961–67	1968–73	1974–77
Percent of strikes over working conditions	16	22	30	35
Percent of wildcat strikes	[a]	32	40	43
Percent of wildcat over working conditions	[a]	59	66	77

Source: Bureau of Labor Statistics, data tape.
[a] Not available.

workers out on strike by making impossible demands. In this case workers' actions are militant, but defensive, and so would be distinct from other aggressive actions.

This analysis would lead us to expect to be able to distinguish four dimensions of industrial conflict: individual actions, and three categories of collective actions—militance within the boundaries of the truce, militance outside the truce, and defensive actions.

In addition to these hypotheses about distinctions among different types of labor militance, I would further propose that these different dimensions would be expected to behave differently over time. Each dimension would have its own degree of cyclical sensitivity. Individual actions would be especially sensitive to labor market conditions, that is, the probability of finding alternative employment if dismissed; collective actions would also be sensitive to the cycle to some extent, since sustained high employment enables unions to replenish strike funds, and workers to pay off debts and accumulate savings. Defensive actions would move countercyclically as employers push workers out on strike when business is slow in order to suspend wage payments without increasing unemployment-compensation liability and to undercut worker militance.

Actions outside the truce would be expected to increase during the 1960's into the 1970's. And defensive actions would be expected to be high when unemployment is secularly high, including during much of the 1950's and the 1970's.

Finally I hypothesize that these different dimensions of industrial conflict would have different effects on the rate of growth of productivity. In general, labor militance would be expected to inhibit productivity growth by interrupting production and therefore output, or by constraining companies from implementing technical changes or reorganizations of the workplace without workers' tacit or formal consent. Those dimensions of industrial conflict with the greatest cyclical variability (individual actions, collective actions to a lesser extent) would play the greatest role in explaining the end-of-expansion slowdown in productivity growth.

The increase in militance outside the truce during the 1960's and 1970's would be expected to help explain the secular decline in productivity growth in the same period. Because actions outside the truce are the least predictable, the most disruptive of production, and the most targeted to working conditions and the production process, such actions should also have a greater impact on productivity growth than militance within the truce.

II. Empirical Investigation

This empirical study should be understood to be preliminary, and hence fraught with all the problems of an exploratory investigation. While in what follows proxies for a number of dimensions of labor militance are developed, they should in no sense be taken as definitive. Strikes are complex events, not uni-dimensional quantities.

Table 2—Results of Factor Analysis of Industrial Conflict, Mining and Manufacturing, 1953–77

	Oblique Factor-Structure Matrices								
	1953–77				1960–77				
Variable	Commu-nality	F1	F2	F3	F1	F2	F3	F4	Commu-nality
Strikes over all issues									
#	.93	.90	−.31	−.17	.89	−.53	−.28	−.24	.85
%Labor force striking	.33	.17	−.52	−.29	.27	−.76	−.33	−.06	.66
%Working days lost	.36	.07	−.55	−.31	.22	−.69	−.33	−.34	.54
Strikes over working conditions									
%All strikes	.77	.81	.33	−.11	.92	.05	−.06	.05	.88
%All strikers	.72	.36	.77	−.12	.23	.26	.06	.77	.70
%Days lost, all strikes	.37	−.02	.59	−.04	.23	.15	.08	.78	.64
Strikes over job security									
%All strikes	.66	−.74	−.18	.47	−.63	.27	.53	.24	.60
%All strikers	.86	−.49	−.01	.89	−.15	.26	.83	−.04	.71
%Days lost, all strikes	.58	−.10	.18	.75	.03	.15	.83	.02	.73
Quit rate	.39	.56	−.19	−.38	.41	−.27	−.52	−.24	.41
Strikes during term of contract									
%All strikes					.89	.10	−.05	.06	.84
%All strikers					.54	.71	.00	.13	.87
%Days lost, all strikes					−.05	.64	.17	.35	.44
Absenteeism rate					.88	−.35	−.34	−.31	.89
Percent of variance explained		33.1	23.4	13.8	34.2	21.1	13.1	9.6	

Source: All variables except quit rate and absenteeism rate, *BLS* "Historical Work Stoppages, 1953–1977" data tape, seasonally adjusted (*SA*). Quit rate for manufacturing, *BLS Employment and Earnings*, *SA*. Absenteeism rate for private nonfarm business, *Current Population Survey*, adjusted for an upward bias due to changes in survey techniques after 1966, series interpolated from annual data.

Therefore all of the measures developed below, in addition to being subject to measurement error, have to be interpreted as indicators of qualitative changes which have not, and often cannot, be precisely quantified.

In order to distinguish among different dimensions of labor militance, I constructed the fourteen variables listed in Table 2. I argue that these variables do not, however, represent separate phenomena but are rather expressions of the four categories of militance just defined: individual actions, collective militance indicating relative labor weakness, collective militance within the truce, and collective militance outside the truce. I therefore performed a factor analysis on the ten variables available from 1953 to 1977, and one on the fourteen variables available from 1960 to 1977, in order to search for underlying patterns of variation.

The results of the factor analyses are presented in Table 2. The first factor for 1953–77 seems best to capture the concept of the unraveling of the truce. Its factor-score variable, denoted *UNTRUCE*, is highly correlated with the number of strikes and the proportion of strikes over working conditions, as well as with quit rates. It is very similar to the first factor produced by the factor analysis on the data available from 1960 to 1977, which is highly correlated with wildcat strikes and absenteeism as well.

The third factor for the 1953–77 data picks up those strikes which are most likely to represent defensive actions by workers and therefore an undermining of their strength rather than labor aggression—strikes over job security. The quarterly scores on this factor will therefore be called *DEFENSE*.

The second factor for 1953–77 loads heavily on the two indicators of labor militance most frequently studied by labor analysts: the number of strikers as a propor-

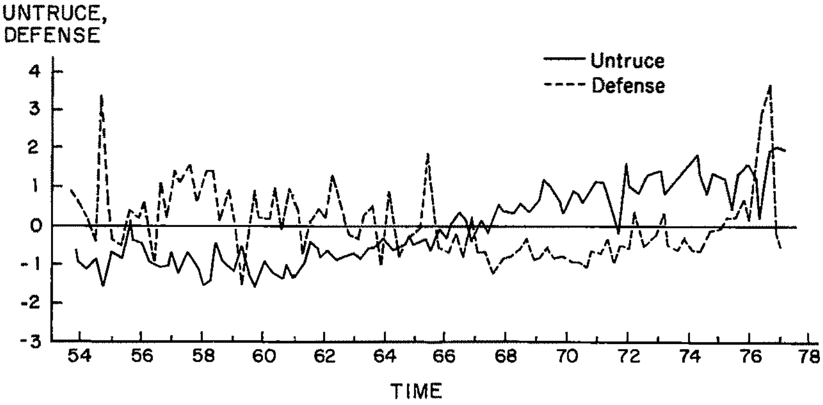


FIGURE 1

tion of the labor force and the proportion of working time lost to strikes. Because the first and third factors load heavily on strikes over nonwage issues, it seems likely that this factor reflects the pattern of strikes over wages and benefits. Such issues are within the bounds of the truce, and the quarterly scores on this factor will therefore be called *TRADITIONAL*.

Thus the factor analysis did provisionally provide some support for two of the dimensions of collective militance hypothesized: it produced an indicator which seems to capture militance outside the truce, *UNTRUCE*, and one reflecting defensive actions by labor, *DEFENSE*. Further definition of the variables highly correlated with *TRADITIONAL* would be required before a satisfactory measure of militance within the truce could be developed.

As yet I have only begun to investigate the question of how these distinct dimensions of labor militance behave over time. By plotting each of the two indicators of different aspects of industrial conflict identified it was possible to further confirm the different behavioral patterns hypothesized for them.

As can be seen from Figure 1, while *UNTRUCE* was low throughout the 1950's, it rose fairly steadily during the 1960's and remained high in the 1970's, as was expected. *DEFENSE* also followed the predicted pattern: it was high and countercyclical in the 1950's, low in the late 1960's and early 1970's, and beginning to increase in the mid-1970's.

In terms of the effect of each of these dimensions of labor militance on productivity growth, to date I have examined simple correlations and regressions. Both *UNTRUCE* and *DEFENSE* have the expected signs, negative for the former and positive for the latter. Furthermore, *UNTRUCE* had even greater explanatory power when the cyclical variation in productivity growth was controlled for.

III. Conclusion

The results of the factor analysis support the proposition that there are different dimensions to labor militance, and that they behave differently over time. Further research is needed on the extent to which they are sensitive to changes in unemployment, and the extent to which they can explain the behavior of productivity growth in the postwar period.

Such research is premised on the expectation, generated by the foregoing analysis, that labor militance matters. If employees are dissatisfied with their working conditions and voice their concerns in ways which challenge managerial initiatives in the workplace, or merely register their preferences by quitting, productivity growth would be expected to fall. If economic hard times enable employers to put labor on the defensive, then *ceteris paribus* productivity growth may accelerate.

The critical implication for policymakers is that any effort to restore growth which does not take into account the distributive impact of increasing productivity will have only a limited success. Instead of simply exacerbating existing antagonisms between workers and companies, policymakers need to address the source of the conflict: the effect of increasing productivity on the quality of worklife.

REFERENCES

Stanley Aronowitz, *False Promises,* New York 1973.

O. Ashenfelter and G. E. Johnson, "Bargaining Theory, Unions and Strike Activity," *Amer. Econ. Rev.,* Mar. 1969, *59,* 35–49.

R. Boddy and J. R. Crotty, "Class Conflict and Macro-Policy: The Political Business Cycle," *Rev. Radical Polit. Econ.,* Spring 1975, *7,* 1–19.

S. Bowles and H. Gintis, "The Crisis of Capital and the Crisis of Liberal Democracy: The Case of the United States," unpublished manuscript, Univ. Massachusetts, Amherst 1979.

R. Freeman and J. Medoff, "The Two Faces of Unionism," *Public Int.,* Fall 1979, *59,* 69–93.

M. Gart, "Labor's Rebellious Rank and File," *Fortune,* Nov. 1966, 150–53.

R. J. Gordon, "The 'End-of-Expansion' Phenomenon in Short-Run Productivity Behavior," *Brooking Papers,* Washington 1979, *2,* 447–60.

John H. Pencavel, *An Analysis of the Quit Rate in American Manufacturing Industry,* Princeton 1970.

W. Simkin, "Refusals to Ratify Contracts," *Labor Relations Reporter,* Nov. 16, 1967.

Federal Mediation and Conciliation Service, *30th Annual Report,* Washington 1977.

U.S.Bureau of the Census, *Statistical Abstract of the United States,* Washington 1951–79.

U.S. Bureau of Labor Statistics, "Historical Work Stoppages, 1953–1977," data tape.

______, *Employment and Earnings,* various years.

A Conflict Theory Approach to Inflation in the Postwar U.S. Economy

By SAM ROSENBERG AND THOMAS E. WEISSKOPF*

Economists have advanced a variety of different explanations for the acceleration of inflation in the *U.S.* economy since the mid-1960's. Some focus on the growth of the money supply; some emphasize tightness in product and labor markets; some stress exogenous shocks affecting price expectations; and others point to labor and/or business market power. Marxian political economists differ among themselves, but many utilize a "conflict theory" framework to explain inflation in contemporary capitalist economies (see Pat Devine and R. E. Rowthorn). Since the conflict between workers and capitalists is central to a Marxian analysis of capitalism, it is only natural for Marxists to model inflation within a framework that emphasizes class conflict over the production and distribution of income. Our purpose in this paper is to develop such a framework for the study of inflation in the postwar *U.S.* economy and to begin an empirical analysis of *U.S.* inflation from the perspective of that theory.

I. A Conflict Theory Model

The conflict theory of inflation focuses on the relationship between *ex ante* claims to income by different classes of people and the income available to satisfy these claims; inflation is linked analytically to an excess of claims over available income. *Ex ante* real claims depend upon the aspirations of the claimants and their ability to act upon their aspirations. Historical, sociological, and demographic factors affect people's aspirations, while political factors, in addition to purely economic ones, affect the ability of classes to put their aspirations into effect. The real income available to meet real claims is the actual gross national product/income expressed in terms of constant dollars of purchased goods and services.

We distinguish initially four broadly defined classes: *workers* concerned with after-tax wage or salary income, *capitalists* concerned with after-tax profits, *transfer recipients* concerned with government transfer payments, and *beneficiaries of government programs* other than transfers. Each of these classes has a common form in which income is claimed and some sense of competition with the claims of the other classes. Each class formulates a real income aspiration to be achieved during a given year; depending upon its ability to act on that aspiration, it presses an *ex ante* claim on real income. Thus, we express total claims (in real terms) as

$$(1)\qquad \overline{W}^c+\overline{R}^c+\overline{G}^c+\overline{\Pi}^c=\overline{Y}^c$$

where $\overline{W}^c$ = total real wage claims (after taxes), $\overline{R}^c$ = total real government transfer claims, $\overline{G}^c$ = total real government program claims, $\overline{\Pi}^c$ = total real (gross) profit claims (after taxes), and $\overline{Y}^c$ = total real claims.

Let $\overline{Y}$ denote the aggregate gross national income available (in real purchasing power) to meet the claims generated in the society. The ratio of total claims to total available income (in real terms) is $\gamma=\overline{Y}^c/\overline{Y}$. Since the available real income is somehow divided among all the competing claimants, *ex post* real receipts must equal *ex post* real available income. Thus,

$$(2)\qquad \overline{W}+\overline{R}+\overline{G}+\overline{\Pi}=\overline{Y}$$

where the *ex ante* claim variables become *ex post* receipt variables without the *c* superscript.

*University of California-Davis, and University of Michigan, respectively. We wish to thank Jack Fulcher and Gerald Serlin for their excellent research assistance.

If $\gamma \gtreqless 1, \bar{Y}^c \gtreqless \bar{Y}$. What is the adjustment process which changes the *ex ante* claims into *ex post* receipts? The key to our model is that the *ex ante* real claims are actually pressed in *money* terms; the claimants multiply their *ex ante* real claims by the anticipated level of prices (for their form of income) in order to establish a money claim that is the same *ex ante* and *ex post*. If γ differs from 1, then the sum of the money claims will differ correspondingly from the available income valued in money terms at anticipated prices. If prices change at a rate different from the anticipated rate, the money claims of the claimants remain the same but the money value of net income available changes. Since real income available $(\bar{Y})$ is given *ex post*, the only possible adjustment to bring about *ex post* equilibrium between $\bar{W}+\bar{R}+\bar{G}+\bar{\Pi}$ and $\bar{Y}$ is an unanticipated change in prices. If $\gamma > 1$, prices will rise more than anticipated, and if $\gamma < 1$, prices will rise less than anticipated. We assume that the supply of money adjusts passively to accommodate price increases rather than posing an independent constraint on price inflation, for the conflict theory approach treats government policy in general and the money supply in particular as endogenous to the forces being modelled.

Thus, in our model, *unanticipated* inflation in any year is the direct consequence of an imbalance between *ex ante* income claims and *ex post* income availability. Although this conflict model directly explains only unanticipated inflation, it can also explain overall inflation insofar as the anticipated component is itself based upon past price behavior and hence on past unanticipated inflation. The key to the usefulness of the model is its ability to explain income claims and income availability in terms of identifiable empirical regularities and/or econometric equations with independent variables reflecting the historical, sociological, and political factors that affect people's aspirations and ability to act on them.

Space limitations prevent us from carrying out here an extensive theoretical analysis of the forces operating on income claims and income availability; we will therefore limit our discussion to a few illustrative points and refer the reader to our earlier paper for a comprehensive treatment of the issues. The starting point for determining the real income available in a given year is the nominal gross national income (Y) generated in the production of gross national output (Q). To deflate Y we do not use the standard *GNP product* deflator (P_q); instead, we utilize an *income* (or purchasing-power) deflator (P_y), which is a weighted average of price indices for consumption, investment, and government purchases. The total gross real income available $(\bar{Y})$ is therefore defined as Y/P_y.

The gross real income available per (full-time equivalent) worker is $\bar{y}$, with lower case symbols denoting per worker values:

$$(3) \quad \bar{y}=\bar{Y}/E=Y/P_yE=P_q\bar{Q}/P_yE=P_{qy}\cdot\bar{q}$$

where E is the number of full-time equivalent workers, P_{qy} is an index of the purchasing power of the product, and $\bar{q}$ is real output per (full-time equivalent) worker. Changes in $\bar{y}$ can thus be analyzed in terms of changes in P_{qy} and $\bar{q}$. The index P_{qy} is primarily a function of the terms of trade; thus a decline in the terms of trade of a given percent will reduce the purchasing power of the product by a percent that is smaller by a factor roughly proportional to the share of international trade in *GNP*. Both P_{qy} and $\bar{q}$ are subject to the forces emphasized by the conflict framework. P_{qy} is affected by distributional conflict on an international scale; to analyze it thoroughly would require the formulation of a conflict model at the level of the world economy as a whole. Productivity $(\bar{q})$ is influenced by a variety of factors, including worker-capitalist conflict over the conditions of work (see David Gordon).

Corresponding to equation (3), we derive an equation for income claims per worker by dividing both sides of equation (1) by E:

$$(4) \qquad \bar{w}^c+\bar{r}^c+\bar{g}^c+\bar{\pi}^c=\bar{y}^c$$

To illustrate our method of analyzing income claims, we will discuss below the forces affecting the real after-tax claims per worker $(\bar{w}^c)$.

We begin by expressing $\bar{w}^c$ as

$$\bar{w}^c = \eta_w \cdot \bar{w}^d \tag{5}$$

where $\bar{w}^d$ is the aspired real after-tax wage per worker, and η_w is a coefficient reflecting workers' ability to act on their aspirations. If $\eta_w = 1$, workers are able to translate their aspirations fully into an *ex ante* claim; to the extent that $\eta_w < 1$, workers are unable to gain a money wage that would meet their desired after-tax real wage at the anticipated level of consumer-good prices. The *ex post* after-tax real wage per worker is affected both by the factors affecting $\bar{w}^c$ and by the behavior of actual consumer-good prices relative to anticipated consumer-good prices; workers' aspirations can be frustrated by (positive) unanticipated inflation as well as by ineffectiveness ($\eta_w < 1$).

We hypothesize that the most important determinant of workers' real wage aspirations ($\bar{w}^d$) is their experience of real wage growth in past years. If real wages have been increasing at a certain annual rate, workers come to expect that it is reasonable and fair for them to continue to enjoy commensurate annual gains. Both recent experience and longer-run trends in real wage growth are likely to have an influence on aspirations for the future. The real wage aspirations of any group of workers are also likely to be related positively to their average skill level, negatively to the average quality of working conditions, and negatively to any trend in their income position relative to the wages of other workers or to the profits of their own employers.

Workers' effectiveness (η_w) in boosting money wages to meet their aspirations will depend on their own strength and on the ability and desire of capitalists to resist their demands. The strength of organized workers depends on the economic and political power of their unions; the strength of unorganized workers depends on the availability of a range of job alternatives. Apart from the impact of union economic power on the bargaining process, union political power is important in a long-run sense insofar as it influences the legal environment for collective bargaining and the government's macro policy.

The economic strength of both union and nonunion workers is likely to vary with current and recent labor market conditions. Tight labor markets increase the likelihood of quitting and decrease the availability of replacements, thereby making capitalists more responsive to worker demands. Also, union strikes are likely to be more successful because there are fewer strikebreakers available and greater alternative employment possibilities to provide additional sources of income for strikers or other members of their families. Tight labor markets in the recent past reinforce the impact of currently tight labor markets by enabling unions to build strike funds and workers to build savings.

Capitalists are more likely to accede to workers' demands for increased money wages the more easily they expect to be able to pass on those increases in higher output prices without suffering losses in sales volume. Their ability to do so depends on such factors as the extent of international competition, the rate of inflation in foreign competitor countries, and the pressures on the federal government to apply monetary restraint. Finally, η_w would obviously be affected negatively by the imposition of wage-price controls in any effective form.

II. Some Empirical Evidence

We will limit our discussion of the evidence here to the presentation of relatively aggregated data in an accounting framework that illustrates the nature of the conflict model and the possibility of its empirical implementation and testing with data from the *U.S.* economy. For more detailed empirical evidence and an explanation of data sources, see our earlier paper.

Table 1 presents some relevant data on inflation, the growth of income availability, and the growth of income claims. For clarity of exposition, we have chosen to summarize annual time-series data in the form of average annual rates of growth during three distinct periods of time: 1954–65, TABLE 1[a]

	1954–65	1965–73	1973–79
Rates of Inflation			
GNI deflator: actual	2.0	4.6	8.0
GNI deflator: unanticipated	0.9	1.6	2.9
Growth of Income Availability			
Real *GNP* per worker	2.7	1.5	0.4
Real *GNI* per worker	2.8	1.4	0.2
Adjusted real *GNI* per worker	2.7	1.3	0.1
Growth of Income Claims Per Worker			
Real after-tax wages (51% of total claims)[b]	3.4	3.4	2.7
Real transfer receipts (7% of total claims)[b]	6.9	9.9	6.6
Real government beneficial programs (11% of total claims)[b]	5.8	2.5	2.7
Real government maintenance programs (11% of total claims)[b]	0.1	−0.5	1.2
Real after-tax net profits (11% of total claims)[b]	6.2	−0.6	1.3
Real depreciation (9% of total claims)[b]	3.6	4.0	4.2
Total real income claims per worker	3.7	2.9	2.9

[a] All figures in average annual percent rates of growth.
[b] Averaged over all three periods.

1965–73, and 1973–79. Any such choice of time periods is bound to be somewhat arbitrary, but we believe that our choice is reasonable as a way of distinguishing three different phases of the inflationary experience of the postwar *U.S.* economy, once it had recovered from the disturbances associated with World War II and the Korean War.

The first part of Table 1 displays the acceleration of inflation in the postwar *U.S.* economy. The average annual rate of increase in the purchasing-power deflator for gross national income (P_y) roughly doubled from each period to the next. The same was true of the unanticipated component of P_y; price expectations have evidently not kept pace with inflationary pressures.[1]

The second part of Table 1 provides salient information on the growth of aggregate income availability, expressed in real terms (based on constant 1972 prices) and per (full-time equivalent) worker. Real *GNP* per worker ($\bar{q}$) grew at successively lower average rates through the three periods. Moreover, because the *U.S.* terms of trade improved slightly in the first period, fell slightly in the second period, and deteriorated significantly in the third period, the average rate of growth of real *GNI* per worker ($\bar{y}=P_{qy}\bar{q}$) dropped even more sharply from period to period. Some minor adjustments required to obtain a fully consistent measure of real *GNI* per worker had a negligible impact on the figures in Table 1, reducing the rate of growth of $\bar{y}$ by one-tenth of a percentage point in each period. Thus we find that the average annual rate of growth of real income availability per worker in the *U.S.* economy was a healthy 2.7 percent from 1954 to 1965; it dropped to one half that rate from 1965 to 1973, and it was reduced virtually to zero from 1973 to 1979.

We turn now to the other side of the balance to investigate the growth of claims

[1] The unanticipated component of inflation was calculated by subtracting from the actual rate an estimate of expected inflation based on the twelve-month Livingston survey forecasts, as presented in John Carlson, and updated in a letter from Carlson.

on real income in the *U.S.* economy, also on a per worker basis. The last part of Table 1 presents the average annual rates of growth of six (exhaustive) categories of claims:[2] after-tax wages ($\bar{w}^c$); transfers ($\bar{r}^c$); government beneficial programs ($\bar{g}_b^c$); government maintenance programs ($\bar{g}_m^c$); after-tax net profits ($\bar{\pi}_n^c$); and depreciation ($\bar{\pi}_d^c$). Government programs include all federal, state, and local government purchases of goods and services; purchases under the budget headings of national defense, international affairs, general administration, and civilian safety were allocated to maintenance programs, while all other purchases were considered beneficial. Depreciation was measured by (adjusted) capital consumption allowances and subtracted from gross profits to determine net profits. The last line of Table 1 shows average rates of growth of total real income claims per worker; each such overall growth rate is a weighted average of growth rates of the six component claim categories, with the weights reflecting the share of each component in total claims during the relevant period.

After-tax wages constitute roughly one-half of the overall claims on total real income; it is noteworthy that after-tax real wage claims per worker rose just as rapidly from 1965 to 1973 as from 1954 to 1965 and then continued to rise, a little more slowly, from 1973 to 1979. Real transfer claims per worker rose more rapidly than any other type of claim in all three periods. Real depreciation claims per worker rose somewhat less rapidly, but very steadily. The remaining claims displayed variable patterns of growth, but in each case grew more rapidly from 1973 to 1979 than from 1965 to 1973.

Total real income claims per worker increased at an annual average rate of 3.7 percent during the first period and then slowed down only moderately to 2.9 percent in both the second and third periods. Thus the growth of income claims did not decelerate at all as sharply as the growth of income available; the difference between the rate of growth of claims and the rate of growth of available income increased from 0.9 to 1.6 to 2.8 percent in the three successive periods. These figures are precisely equal (except for rounding error) to the rate of unanticipated inflation.

III. Conclusion

Our objectives in this paper have been relatively limited. We sought to elaborate upon the theoretical framework of a conflict theory of inflation and to begin to organize data on the *U.S.* economy along lines suggested by that framework. The data we have compiled show clearly that the ability of the *U.S.* economy to generate growth in per worker real income has greatly diminished since the mid-1960's. Moreover, we have developed plausible estimates of claims on real income which indicate clearly that there has not been a commensurate slowdown in the rate of growth of claims placed upon the economy. These data provide empirical substance to a conflict theory interpretation of postwar inflation in the *U.S.* economy, whereby (unanticipated) inflation is explained in terms of the difference between the growth of real income claims and the growth of real income availability.

It should be stressed that thus far we have merely developed a useful accounting framework for illustrating empirically the conflict theory of inflation. Our procedures for measuring unanticipated inflation and for estimating *ex ante* income claims are linked in such a way as to assure tautologically that unanticipated inflation will be equal to the difference between the rate of growth of real income claims and real income availability. The real test of the usefulness of the conflict theory approach lies in whether or not it can help to explain (and hence also help to predict) the rate of growth of various types of real income claims and the rate of growth of real income availability.

The full implementation of our model will require extensive disaggregation of the *U.S.*

[2] The figures in the last part of Table 1 represent averages of the percentage increase of a given year's *ex ante* claim over the previous year's *ex post* receipt. *Ex ante* real claims were calculated by dividing *ex post* money values by the anticipated level of the relevant price index; anticipated price levels were calculated by means of the estimates cited in fn. 1.

economy and further subdivision of the claimant classes we have discussed here. Once an appropriate level of disaggregation has been reached, we will formulate precise hypotheses to explain income claims and income availability and test them rigorously with econometric techniques.

For the present, we draw some encouragement about the usefulness of the conflict theory approach from the aggregate data we have compiled. The fact that the growth of real wage claims has been relatively steady over the past 25 years suggests that previously experienced rates of growth of real income have a significant influence on future claims, and that current economic conditions (which are stressed in most alternative analyses) are relatively less significant. In work that we have already done on the growth of other categories of income claims, reported in our earlier paper, we have found some stable patterns of behavior that seem likely to be explainable in terms of the demographic, historical, sociological, and political factors emphasized by the conflict theory approach. Thus we are reasonably confident that our future research along these lines will provide empirical support for the applicability of a conflict theory of inflation to the postwar *U.S.* economy.

REFERENCES

J. A. Carlson, "A Study of Price Forecasts," *Annals Econ. Soc. Measure.*, Winter 1977, *6*, 27–56.

P. Devine, "Inflation and Marxist Theory," *Marxism Today*, Mar. 1974, *18*, 70–92.

D. M. Gordon, "Capital-Labor Conflict and the Productivity Slowdown," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 30–35.

S. Rosenberg and T. E. Weisskopf, "A Conflict Model of Inflation Applied to the Postwar U.S. Economy," work. paper no. 157, Univ. California-Davis, 1980.

R. E. Rowthorn, "Conflict, Inflation and Money," *Cambridge J. Econ.*, 1977, *1*, 215-39.

Capacity Utilization Measures: Underlying Economic Theory and an Alternative Approach

By Ernst R. Berndt and Catherine J. Morrison*

Measures of industrial capacity utilization (hereafter, *CU*) have been used extensively in helping to explain changes in the rate of investment, labor productivity and inflation. The *CU* measures have also been used to obtain indices of capital in use, as distinct from capital stock in place. A number of alternative measures of *CU* are periodically calculated and published; the 1980 *Economic Report of the President*, for example, contains three series, that by the Federal Reserve Board, the U.S. Department of Commerce (Bureau of Economic Analysis) and the Wharton School of Finance. Other publicly available series are those prepared by McGraw-Hill Publishing Company, the U.S. Department of Commerce (Bureau of the Census), and Rinfret-Boston Associates, Inc.

Although a host of *CU* measures is publicly available, it is not at all clear how one should interpret changes over time in each measure or variations among them. A principal reason underlying these interpretation problems is that the crucial link between underlying economic theory and the constructed measure of *CU* is weak.

One way in which this issue has manifested itself in the policy domain over the last five years has been with respect to the uncertain effects of dramatic increases in energy prices on capacity output and on *CU*. Each of the *CU* measures noted above is computed in such a way that explicitly ignores any role for energy prices. Yet several times during the last decade, though growth to apparently high rates of *CU* had taken place, investment and average labor productivity were much lower than expected, and the rate of price increase much greater. In brief, during the last decade the explanatory power of alternative *CU* measures has dropped sharply. Some have conjectured that post-OPEC energy price increases may have brought about major changes in the *U.S.* economy so that old quantitative relationships between measured *CU* and investment, labor productivity, and price inflation may have been altered substantially.

In order to assess effects of changes in P_E on *CU*, a re-examination of the notion and measurement of *CU* is needed, based on the framework of the economic theory of the firm. That is the focus of this paper.

I. Theoretical Foundations

The concepts of capacity output and capacity utilization are inherently short-run notions, conditional on the firm's stock of quasi-fixed inputs. Consider a firm with a production function

$$(1) \qquad Y=f(v,x)$$

where Y is the flow of output, v is an $n\times 1$ vector of variable inputs, and x is a $j\times 1$ vector of service flows from quasi-fixed inputs (inputs fixed in the short run, but available at increasing marginal costs in the long run). As discussed by W. Erwin Diewert, Lawrence Lau and Daniel McFadden, the optimization problem facing the firm is typically characterized as that of maximizing variable profits (revenue minus variable costs), conditional on output price P, prices of the variable inputs P_v, and x. An alternative framework employed in this paper is

*Professor of applied economics, Massachusetts Institute of Technology and Canada Council Fellow, University of British Columbia, respectively. Research support through NSF grant 7910488-DAR is gratefully acknowledged.

based on recent developments in the theory of duality, and puts forth the optimization problem facing the firm as that of minimizing variable costs, conditional on Y, P_v, and x. In this dual approach, given appropriate regularity condition on the production function (1), there exists a dual variable cost function

$$Cv = g(Y, P_v, x) \tag{2}$$

where C_v is average variable cost. Let P_x be the vector of rental prices for the quasi-fixed factors, and define average total cost C as

$$C = C_v + C_f \tag{3}$$

where C_f is average fixed cost.

The definition of capacity output Y^* used here is that level of output for which C is minimized, i.e.,

$$Y^* = h(P_v, x, P_x) \tag{4}$$

This concept of capacity output—that level of output at the minimum point of the short-run average total cost curve—dates back at least to J. M. Cassells, was suggested later by Lawrence Klein, and yet has hardly ever been examined empirically. When there are long-run constant returns to scale, Y^* also represents a tangency between the long-run and the short-run average total cost curves. In the more general case when there are nonconstant returns to scale, one can define Y^* as that level of output at which the short- and long-run average total cost curves are tangent. Hereafter, however, we shall assume long-run constant returns to scale.

It should be noted in passing that this definition of Y^* will generally differ from a capacity output level Y^{**} defined as that level of output maximizing variable profits; generally for a competitive firm, Y^{**} will be greater than (less than) Y^* when the exogenous output price is greater than (less than) the minimum level of short-run average total costs.

Now define the rate of capacity utilization u as actual output Y over capacity output Y^*, i.e, $u = Y/Y^*$. Until recently, empirical efforts to measure Y^* and u have been hampered by the lack of appropriately flexible functional forms and, more importantly, by basic developments in the underlying economic theory. Such limitations were noted by Klein. Theoretical contributions by Diewert, Lau, and McFadden, however, have enriched the range of empirical research possibilities substantially, and in particular now enable one to obtain econometric estimates of Y^* and u. Some recent estimates will be discussed later in this paper.

Given a clear notion of economic capacity output and capacity utilization, we now address the issue of how variations in input prices might affect Y^* and u. Assume for the moment that there is only one quasi-fixed factor, physical capital (K). An important issue concerns how a change in the price of a variable input such as energy (E) might affect Y^*. In particular, does an increase in P_E shift the minimum point of the short-run average total cost curve to the right (increasing Y^*), to the left (decreasing Y^*), or does it merely shift the average cost curve upward without affecting Y^*? To the best of our knowledge, no published theoretical research has been done concerning such an issue.

In an unpublished paper, Robert Rasche and John Tatom (1977b) show that if the variable input (in this case, E) and the fixed input K are Hicks-Allen substitutes (complements), then an increase in P_E decreases (increases) Y^*; if, however, E and K are independent inputs such that long-run substitution elasticities between E and K are zero, then variations in P_E do not affect Y^*. Intuitively, this phenomenon can be explained as follows. If E and K were substitutable inputs, with an increase in P_E the firm's long-run optimal K/Y ratio would increase from, say, K_0/Y^* to K_1/Y^*. This implies that, in the short run, the given level of capital K_0 corresponds with a smaller Y^*. Alternatively, if E and K were complementary inputs, an increase in P_E would imply a lower long-run optimal K/Y^* ratio; in the short run, the firm's given level of capital K_0 would then be associated with a large Y^*. Obviously, matters become more complicated when there are multiple quasi-fixed factors and/or multiple outputs.

In their empirical research, Rasche and Tatom (1977a) assumed a Cobb-Douglas function which implies the assumption of *K-E* substitutability. Conditional on this Cobb-Douglas assumption, Rasche and Tatom concluded that increases in P_E since 1973 have reduced the nation's Y^* considerably—by about 10 percent—and that therefore any expansionary monetary or fiscal policy would be ill-advised. Even though published measures of u might indicate some slack capacity, according to Rasche and Tatom the true economic level of u is considerably higher than the published measures would indicate. Results would have been different, of course, had they not assumed *K-E* substitutability. Space considerations preclude our discussing the empirical evidence here on *E-K* substitutability vs. *E-K* complementarity; for a recent review, see Berndt and David Wood. What is clear, however, is that the typical published measures of capacity utilization are not very informative in terms of assessing the economic effects of increases in P_E.

Another example illustrating the importance of Y^* and u is the effect of increases in P_E on the recent productivity slowdown in the United States and other industrialized countries. It has long been known that both multifactor productivity and average labor productivity tend to be procyclical. What has surprised economic observers in recent years is that productivity trends have not followed the traditional procyclical patterns, and have shown very little growth in spite of apparently high and at times increasing levels of *CU*. Since the published measures of *CU* might be unreliable, comparison of recent productivity trends with those of earlier periods could be misleading if post-1973 P_E variations had affected Y^* and u. Clearly, what is needed is additional theoretical and empirical research on the notion and measurement of Y^* and u.

II. Recent Estimates of an Economic Measure of *CU*

Earlier it was noted that developments in the theory of duality now permit estimation of Y^* and u with general functional forms. Three recent empirical studies in this vein are those of Berndt (1980), Morrison, and of Berndt, Morrison, and G. Campbell Watkins. Based on earlier work by Berndt, Melvyn Fuss, and Leonard Waverman, the dynamic optimization problem facing the firm is specified as that of minimizing the present value of costs, given Y, P_v, P_x, increasing marginal internal costs of adjustment for x, and positive initial levels of the quasi-fixed inputs.

In Table 1 we reproduce two alternative measures of economic capacity utilization for *U.S.* manufacturing, 1958–77; both are taken from Berndt (1980, Table 6), where additional details are provided. The first measure is based on a dynamic cost function model with a single quasi-fixed factor K. The other measure incorporates an observation by Walter Oi that certain types of skilled labor should also be considered as quasi-fixed factors; hence the second measure is based on a dynamic model with two quasi-fixed factors, K and W, where W is hours at work of nonproduction (white collar) workers in *U.S.* manufacturing. For purposes of comparison, we also reproduce in Table 1 the Wharton and Federal Reserve Board (FRB) measures, as well as estimated ratios of short-run marginal cost to long-run average total cost, evaluated at the actual level of output. These measures assume that input supply curves are perfectly elastic, and represent the cost inflationary consequences of producing at output levels different from Y^*.

Several comments are in order. First, the economic measures are always greater than unity, whereas the Wharton and FRB figures are always less than unity. To some extent this can be interpreted as merely a scaling convention, since Wharton and FRB measures approaching 90 percent are typically viewed as signifying very near "full capacity." On the other hand, that the economic measures are greater than unity is informative, for it indicates that production is to the right of the minimum point of the short-run average total cost curve, thereby inducing cost-reducing net investment.

Second, simple correlations among the various measures indicate considerable dif-

TABLE 1—ALTERNATIVE MEASURES OF CAPACITY UTILIZATION, AND RATIO OF ESTIMATED SHORT-RUN MARGINAL COST TO LONG-RUN AVERAGE TOTAL COST *U.S.* MANUFACTURING, 1958–77

	Capacity Utilization Model with:				*SMRC/LRAC* Model with:	
Year	*K* Fixed	*W*, *K* Fixed	FRB Measure	Wharton Measure	*K* Fixed	*W*, *K* Fixed
1958	1.106	1.091	0.752	0.742	1.015	1.026
1959	1.110	1.118	0.819	0.789	1.013	1.030
1960	1.171	1.131	0.802	0.769	1.024	1.035
1961	1.177	1.130	0.774	0.737	1.026	1.036
1962	1.197	1.145	0.816	0.765	1.027	1.038
1963	1.224	1.167	0.835	0.777	1.031	1.044
1964	1.226	1.164	0.856	0.795	1.031	1.042
1965	1.232	1.190	0.896	0.842	1.030	1.045
1966	1.214	1.170	0.911	0.882	1.027	1.040
1967	1.184	1.129	0.869	0.869	1.026	1.033
1968	1.178	1.119	0.871	0.892	1.024	1.030
1969	1.169	1.108	0.862	0.902	1.025	1.029
1970	1.111	1.026	0.793	0.841	1.018	1.012
1971	1.110	1.052	0.784	0.827	1.017	1.015
1972	1.204	1.139	0.835	0.879	1.033	1.037
1973	1.240	1.185	0.876	0.932	1.040	1.049
1974	1.092	1.079	0.838	0.905	1.012	1.018
1975	1.160	1.096	0.729	0.798	1.030	1.026
1976	1.259	1.170	0.795	0.860	1.051	1.048
1977	1.267	1.183	0.819	0.887	1.055	1.052
Mean	1.182	1.130	0.827	0.834	1.028	1.034

ferences. For the single (two) quasi-fixed factor model, simple correlations between the economic measure and the FRB index are .419 (.523), while those between the economic measure and the Wharton index are only .244 (.140). The simple correlation between the Wharton and FRB index is .605. Both economic measures of *CU* indicate relative peak years in 1965, 1973, and 1977, while peak years for the FRB index are 1966, 1973, and 1977. The Wharton relative peaks are in 1966, 1969, 1973, and 1977. Economic capacity utilization measures are lowest in 1958–59, 1970–71 and 1974–75, essentially coinciding with low points of the Wharton and FRB measures, although the latter both indicate slight downturns in 1961.

The economic capacity utilization measures differ from the Wharton and FRB values in one very important respect, however. According to the FRB measure, the relative peak years of 1973 (.876) and 1977 (.819) were considerably smaller than the 1966 all-time peak (.911), whereas, for the economic measures the 1973 and 1977 peaks are virtually identical. The Wharton index differs slightly; its all-time peak is 1973 (.932), and the 1966 (.822) and 1977 (.887) peaks are about equal but smaller than that in 1973. One implication of these results is that if one believes these measures of economic *CU*, then the much heralded 1973–77 productivity slowdown relative to 1965–73 cannot be attributed to the end year 1977 being of lower capacity utilization.

Finally, although not reported in Table 1, we have calculated the effects of increased P_E on Y^* in *U.S.* manufacturing. For 1977 the estimated elasticity of Y^* with respect to P_E is positive (due to *E-K* complementarity), but quite small. Berndt (1980) reports a 1977 estimate of this elasticity as 0.021 for the only *K*-fixed model, and 0.047 for the model with both *W* and *K* fixed. Hence these figures suggest that although energy price increases have affected Y^* in *U.S.* manufacturing, the quantitative magnitude is modest. These small positive estimates contrast sharply with those of Rasche and Tatom (1977a), who used a Cobb-Douglas model and estimated that the elasticity of Y^* with respect to P_E was about $-.10$.

III. Concluding Remarks

We have argued that greater attention should be focussed on developing better notions and measures of *CU*; we have also reproduced several recent economic measures of *CU*, and have shown that they often differ considerably from the more familiar FRB and Wharton measures. The Wharton index, it will be recalled, is essentially a ratio of actual *Y* to potential output, where the latter is based on previous peak values of the output-capital ratio and cumulative net investment. Factor prices and quantities of noncapital inputs are not incorporated. Another measure of *CU* is really an engineering notion of *capital* utilization rather than an economic measure of *capacity* utilization; this is based on the ratio of actual electricity consumption to the maximum possible electricity consumption, where the latter is obtained using the rated horsepower capacity of electric machinery. For an example of such a procedure, see Murray Foss.

Although we are somewhat reluctant to advocate publication of yet another series of *CU*, we hope that applied researchers in the future will devote greater attention and care to the economic theory underlying the concept of capacity output, and will publish series which can then be interpreted more clearly.

REFERENCES

E. R. Berndt, "Energy Price Increases and the Productivity Slowdown in United States Manufacturing," in *The Decline in Productivity Growth*, Fed. Reserve Bank Boston *Conference Proceedings*, No. 22, Boston 1980.

______, **M. A. Fuss, and L. Waverman,** "A Dynamic Model of Costs of Adjustment and Interrelated Factor Demands, with an Empirical Application to Energy Demand in U.S. Manufacturing," disc. paper no.79-30, Dept. Econ., Univ. British Columbia, Nov. 1979.

______, **C. J. Morrison, and G. C. Watkins,** "Dynamic Models of Energy Demand: An Assessment and Comparison," in Ernst R. Berndt and B. C. Field, eds., *Measuring and Modelling Natural Resource Substitution*, Cambridge: M.I.T. Press, forthcoming.

______ **and D. O. Wood,** "Engineering and Econometric Interpretations of Energy-Capital Complementarity," *Amer. Econ. Rev.*, June 1979, *69*, 342–54.

J. M. Cassels, "Excess Capacity and Monopolistic Competition," *Quart. J. Econ.*, May 1937, *51*, 426–43.

W. E. Diewert, "Applications of Duality Theory," in Michael D. Intriligator and David A. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland 1974, 106–71.

M. F. Foss, "The Utilization of Capital Equipment: Postwar Compared with Prewar," *Surv. Curr. Bus.*, June 1963, *43*, 8–16.

L. R. Klein, "Some Theoretical Issues in the Measurement of Capacity," *Econometrica*, Apr. 1960, *28*, 272–86.

L. J. Lau, "A Characterization of the Normalized Restricted Profit Function," *J. Econ. Theory*, Feb. 1976, *12*, 161–63.

D. F. McFadden, "Cost, Revenue and Profit Functions," in Melvyn A. Fuss and Daniel F. McFadden, eds., *Production Economics: A Dual Approach to Theory and Applications*, Vol. 1, Amsterdam: North-Holland Publishing Co., 1978, 3–109.

C. J. Morrison, "Investment Decision of Firms with Non-Static Expectations," unpublished paper, Dept. Econ., Univ. British Columbia, June 1980.

W. Y. Oi, "Labor as a Quasi-Fixed Factor," *J. Polit. Econ.*, Dec. 1962, *70*, 538–55.

R. H. Rasche and J. A. Tatom, (1977a) "The effects of the New Energy Regime on Economic Capacity, Production and Prices," *Fed. Res. Bank St. Louis Rev.* May 1977, *59*, 2–12.

______ **and** ______, (1977b) "Firm Capacity and Factor Price Changes," xerolith, Fed. Res. Bank St. Louis, not dated.

Stochastic Equilibrium and Capacity Utilization

By ARTHUR DE VANY AND N. G. FREY*

This paper is a progress report on research into the utilization of capacity in markets with stochastic demand and production relations which use some form of quantity rationing as a clearing mechanism. The work has developed monopoly and competitive models of inflexible price-quantity rationing markets which have been applied to several industries, including the trucking, dental, and steel industries. In this report we briefly state the main theoretical results for the competitive model, emphasizing those which pertain to properties of the steady state. The empirical work is used to interpret the theory and shed light on the economics of slack capacity. We close with some extensions and generalizations which give an indication of where the work is headed.

I. The Stochastic Production Function

When production and demand are random, new considerations are involved in defining a production function. Even with a certain production process, production does not occur unless there is an order to be filled. If demand is random there is only a probability of an order to be filled. This means the production process is switched on and off by arrival of orders and cannot be observed as a pure, steady production process. Even if the firm smooths production through inventory adjustment, this can be considered a random ordering of the firm's output by itself in anticipation of future demand, and does not alter the fundamental point. The production relations used in economic models assume unlimited orders are available for the firm to fill through production. With random demand this assumption is not fulfilled. One must instead consider the rate at which the plant could produce, were an unlimited number of customers present, and the probability a customer or order is in the system so that the process is switched on.

More generally, the production relation is a distribution of possible outputs which may be defined over fixed, nonstochastic inputs. The conditional mean of the production distribution, conditional on there being orders to fill, could serve as a production relation. Let

$$(1) \qquad s = s(K, L \mid q \geqslant 1)$$

denote the mean output rate of the process for fixed capital and labor input vectors K and L, conditional on q = one or more customers in the system. Two useful interpretations of this expected production function are that output is random because there are variations in the efficiency of production due to noise or random inefficiencies, or that the process is efficient and certain, but differences among customers or their orders give rise to different times to fill an order.

II. Demand Specification

Demand is assumed to be random and each agent's demand distribution is a function of price and expected rationing cost. Analyses of alternative forms of quantity rationing indicate there is a class of rationing policies which generate a rationing cost function which is increasing and convex in the firm's expected output, and decreasing and convex in the firm's mean production capability. In this class of rationing, policies are fixed-price policies which ration excess demand by queue, inventory stock outs, or prorationing.

If consumers are assumed to be identical and search to minimize the sum of price and expected rationing cost, which we call the full price, then full-price dispersion collapses and a single market-clearing full price

*Simon Fraser University and California State University-Hayward, respectively.

emerges. In this case, expected demand at the firm level may be written

$$q=f[p+R(q,s)] \tag{2}$$

with the firm subject to the constraint of the market full price. Stability requires $1>f'R_q>0$ and $\partial q/\partial s=-R_s/R_q=-1/R_q$. Full-price taking implies the firm may choose p and s or p and q, but must take the level of the remaining free variable as determined by the market. For fixed full price, demand is downward sloping in price, and increasing capacity shifts demand to the right, increasing price elasticity and output. If expected rationing cost is homogeneous of degree zero in expected output and capacity, then output increases proportionately with capacity.

The competitive firm maximizes expected profits subject to the full-price constraint by choosing any two of price, output or capacity. The expected profit function is

$$\pi=[P-R(q,s)]q-C(q,s) \tag{3}$$

with expected price given by $p=P-R(q,s)$, $P=$ full price and expected cost, $C(.)$, a convex cost function of expected output and capacity. With $R(.)$ and $C(.)$ convex, the profit function is concave and a maximizing solution exists.

III. Stochastic Equilibrium

Customers have a reservation full price P^*. A firm charging price p will find that its upper limit on demand is q^* such that $P^*=p+R(Q^*,s)$. This means the demand distribution is truncated at the upper limit q^*. This condition establishes a finite holding capacity for the system and is sufficient for the existence of a steady-state equilibrium for a fairly broad class of demand and production distributions.

Let $\beta>0$ be the probability demand exceeds q^*. Then expected demand given q^* is $E[q|q^*]=(1-\beta)E[q]$, where $E[q]$, is the unconditional expectation of demand. Now define $l=E[q]/s$ as the expected load on capacity. It can then be shown that the effective expected load on capacity, net of those who are rationed out, is $l'=(1-\beta)l=(1-\beta)(E[q]/s)$. Moreover, if an equilibrium exists, then $l'<1$. This theorem establishes the existence of technical excess capacity, that is, output q' is strictly less than the mean output that could be produced were customers always available.

The output q' is produced by the plant operating at a utilization rate of $l'=(1-\beta)l$. Full utilization can only be achieved if $\beta\equiv 0$, a result which would be true only if rationing cost is zero, that is, if the demand and production distributions were degenerate. Of the customers who arrive at the firm, the proportion β are rationed and search for service at other firms. Each customer who enters the market searches an average $\sigma=1/1-\beta$ times per unit of output. Thus, the number of firm contacts by searchers is $(1/1-\beta)q$, an amount which exceeds demand. We have shown that output is less than demand and demand is less than searches. These quantities become equal as $\beta\to 0$ and production approaches its technical limit as well.

IV. Some Theorems and Tests

THEOREM 1: *Price—The competitive firm's price exceeds expected marginal cost, i.e.,* $p=C_q+qR_q$.

THEOREM 2: *Excess Capacity—The competitive firm's expected output is less than the output level which would minimize average cost.*

This theorem establishes the existence of economic excess capacity whereas above we showed the existence of technical excess capacity. To prove the theorem we first note that, by Theorem 1, competitive price exceeds marginal cost by an amount equal to the marginal rationing cost borne by all q of the firm's customers consequent to an increase in output. But if $p>C_q(q,s)$ and profits are zero, then $p=C(q,s)/q>C_q(q,s)$, which indicates marginal cost is less than average cost.

THEOREM 3: *Uncertainty and Capacity—The quantity rationing expected profit-maximizing competitive firm produces any given output with more capacity than the same firm would employ with certain demand.*

This follows directly from the existence of positive rationing costs, which are internalized to the full-price-taking firm. Since rationing cost is decreasing in capacity, s, the sum of production and rationing cost is minimized at a capacity larger than the level which minimizes production cost alone.

It follows from Theorems 2 and 3 that expected rationing cost is a variable cost of production. The variable factors involved are not owned by the firm and, hence, do not appear in its production and cost relations. Therefore, a price which just equals the marginal cost of firm-owned variable factors will not cover all variable factors, and will be inefficient because it will not support the proper level of capacity nor provide the correct incentives for customers.

The rather startling interpretation which may be placed on this result is that the firm rents variable factors from its clientele by charging a price which is less than the reservation full price by an amount equal to expected rationing cost. An empirical confirmation of the hypothesis that rationing cost is the cost of customer supplied inputs to the production process comes from a study of the dental industry (see De Vany et al.). Arguing that patient waiting time is a variable input in the production of dental care, the study estimated a production function including waiting time as well as more conventional dental inputs. A significantly positive marginal product of time was found and firms were also found to employ the profit-maximizing level of patient waiting time.

A further implication of these theorems involves the so-called public goods aspect of excess capacity. It has been argued, and may even be generally accepted, that if a customer appears at a facility having excess capacity he should pay only marginal cost, since his use of the facility preempts no other user. But it is simply a consequence of making this customer a nonrecurrent unanticipated event which renders the fixed capital sunk with respect to that event and hence nonrecoverable. What we have to do is ask what happens if this customer is a unit increase in long-run demand, in which case he raises the probability the facility will be full and, hence, increases expected rationing cost. In this event price must equal the full marginal cost including marginal production rationing cost. This price will also induce the correct scale decision since it is known to equal the marginal cost of the plant size increase required to hold rationing cost constant when output increases by one unit.

A final implication of these theorems is that competitive prices depart optimally from marginal cost in the manner indicated by the Baumol-Bradford rule. Since there are short-run decreasing costs, price must depart from marginal cost. The price elasticity in the Baumol-Bradford formula implied by full-price taking is $\varepsilon = p/qR_q$. As rationing cost goes to zero $R_q \to 0$ and $\varepsilon \to -\infty$ implying that $P \to C_q$. The quantity rationing competitive model contains the standard competitive model as a limiting case.

THEOREM 4: *Price Dispersion—Nondegenerate price dispersion is supportable by diverse consumers.*

If customers were identical, firms could meet the market full price with an infinity of price and waiting time combinations. If firms have identical production functions they will all choose the unique cost-minimizing combination of firm-owned and customer-owned inputs. The payment to identical customers must be equal and so price must be equal among all firms. If firms produce with different input combinations, they employ different amounts of waiting time from their customers, which is rational only if the cost of customer time inputs differs among customers.

Two implications of Theorem 4 have been tested. One asserts that price and rationing cost would be negatively correlated, the other that the ratio of price to customer inputs should be higher the higher the cost of those inputs. These predictions were confirmed in another study of dental practice (see De Vany, Donald House, and Thomas Saving).

THEOREM 5: *Entry—Entry increases market demand and decreases output per firm in a neighborhood of equilibrium.*

The latter part of the theorem must be true if the equilibrium is stable, since profits must fall with entry. The first part is true because if density of firms relative to customers rises, rationing cost falls and quantity demanded must increase.

The theorem explains the heretofore puzzling finding that in medical markets per capita demand is positively related to the density of medical care providers per capita. In De Vany, House, and Saving it is shown that density is a proxy for waiting time and that per capita demand responds to changes in waiting time in the manner predicted by Theorem 5.

THEOREM 6: *Critical Order Limit—For each customer and firm i there exists a critical number q_i^* such that if this number of orders is in process at firm i the customer does not place an order.*

This is a central result of the search process which distributes customers among quantity rationing firms. It is perhaps the most difficult proposition to test because of the difficulty of observing refusals to place an order. In our earlier paper an indirect test was made by examining the behavior of order backlogs in the steel industry. It was found that new orders and shipments obeyed a Poisson distribution, while the number in the order backlog exhibited a negative binomial distribution. A negative binomial distribution arises in an experiment consisting of a sequence of Bernoulli trials and gives the probability of r failures before the nth success. Thus the order backlog distribution is consistent with the hypothesis that some customers refuse to join the order backlog (a failure). Moreover, the distributions estimated allow direct calculation of the probability a randomly arriving order will join the order backlog. This turned out to be a monotone decreasing function of the number in the backlog. Since this holds for industry level data, it supports the proposition that industry effective demand is less than potential demand and that demand is negatively related to rationing cost. We did conclude that the order backlog functions as an implicit futures market in steel productive capacity.

Of particular relevance to the economics of slack capacity, this result allows us to give rough estimates of the elasticity of effective demand with respect to slack capacity. Our work brackets the elasticity in the .10 to .20 range, with a mode of .14. This suggests a 10 percent increase in slack capacity increases output by about 1.4 percent.

V. Directions for Extension and Generalization

A nonoperational finding not referred to is that, under fairly weak conditions, a competitive quantity rationing market is efficient. This finding removes some of the objections economists would have against a market that operates with inflexible prices and short-run quantity rationing. But closer examination reveals that the models we have been describing actually employ a more complex clearing mechanism or exchange contract than a pure competitive auction market. The quantity rationing market employs both price and nonprice rationing instruments to achieve its allocations. The allocations which are supportable with exchange conventions which allow both price and quantity allocations cannot be inferior to those supported by prices alone. Pure price rationing is a special form of exchange convention. In some market settings, efficient allocations may not be supportable with a pure price mechanism, and an extended mechanism may be called for to achieve efficiency. The models described here are one class of mechanisms—those using short-run inflexible prices and quantity limits—there are many other mechanisms that deserve study.

REFERENCES

A. S. De Vany, "Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality," *J. Polit. Econ.*, June 1976, *84*, 823–41.

______**and T. R. Saving,** "Product Quality, Uncertainty and Regulation—The Trucking Industry," *Amer. Econ. Rev.*, Sept. 1977, *67*, 583–94.

______**and N. G. Frey,** "Price Stability Order Backlogs and the Value of Capacity in the Steel Industry," mimeo., Simon Fraser Univ., Nov. 1979.

______**et al.,** "Patient Waiting Time as an Input, Regulation and Production in the Dental Industry," mimeo., Texas A & M Univ., Feb. 1980.

______**, D. R. House, and T. R. Saving,** "The Role of Patient Time and Utilization of Dental Firm Capacity in the Pricing of Dental Services," in Edwin H. Mills, ed., *Competition and Regulation in the Health Care Market*, Chicago: Blue Cross/Blue Shield, forthcoming.

Long-Run Changes in the Workweek of Fixed Capital

By MURRAY F. FOSS*

The measurement of capital has always been difficult and the measurement of fixed capital input in accounting for the long-run growth of output in the United States has had its share of controversy. One question that arose in what could be called the capital utilization dispute of the late 1960's and 1970's was whether, in measuring capital input, the change in the real capital stock in place (the balance-sheet concept) should be adjusted also for any long-run change in hours worked by capital per week or per year. This issue was part of a broader debate over the sources of *U.S.* economic growth. Two key conclusions of growth studies by investigators in the 1950's and early 1960's were 1) the contribution of fixed capital to output growth, while considerable, in a sense has been smaller than commonly thought, and 2) the contribution of the growth in total factor productivity—the "residual"—has been comparatively large (see Moses Abramovitz, Edward Denison, Solomon Fabricant, John Kendrick, and Robert Solow). But this view of the economy was criticized. For example, in 1967 Dale Jorgenson and Zvi Griliches maintained that, as measured by Denison, the contribution of capital was understated (and total factor productivity overstated) in large part because no allowance was made for the long-run increase in the utilization of fixed capital. They made a utilization adjustment that relied on a 1963 study I had made that found an increase in equipment utilization of one-third to one-half from the 1920's to the mid-1950's in the manufacturing sector. My earlier study speculated that a rise in shiftwork, an increase in the importance of continuous industries and more efficient use of fixed capital by business were among the reasons for the increase in hours.

There isn't time to review that controversy (see *Survey of Current Business*) and its aftermath here. The debate came to a halt mainly because the opposing sides acknowledged that data were lacking to make an adjustment of capital for a long-run increase in utilization. It was agreed that it made no difference, given proper measurement, whether one adjusted capital for longer hours, or identified the effect of longer capital hours as a component of the change in total factor productivity. It was further agreed that data on shift work would help eliminate a deficiency that had been pointed out earlier by Abramovitz, Fabricant, and Kuznets. In his 1957 article, Solow speculated that a secular increase in shift work would explain at least part of the 1943–49 deviations from his long-run relationship between output and capital. William Fellner had used a similar explanation for the low capital-output ratio he found for the wartime years.

This paper summarizes major findings of a more detailed study of the change in the workweek of fixed capital in manufacturing from 1929 to 1976. The study, which refers to weekly hours of plant operations and not to the workweek of labor, first sets out the main facts of the change in weekly plant hours in manufacturing from 1929 to 1976, and then analyzes reasons for the changes, using detailed industries as units of observation. The study is scheduled to be published in early 1981.

The data on weekly plant hours are based on a virtually identical set of questions that have been asked in recent years of a sample of manufacturing plants as part of a Census Bureau survey of capacity utilization, and that were asked of all manufacturers in the

*Visiting scholar, American Enterprise Institute for Public Policy Research. The assistance of Robert Dubman is gratefully acknowledged.

1929 Census of Manufactures.[1] The 1929 questions, which are part of the questionnaire reproduced in the published 1929 census volumes, were answered by respondents but apparently were never tabulated. I found that the basic census records for 1929 had been preserved on microfilm, contrary to some earlier information I had received. With the help of the National Science Foundation and the Bureau of the Census, I initiated a study in which the Census Bureau drew from the 1929 returns a probability sample of some 9,000 plants, which was blown up to universe totals for comparison with the results of recent surveys.

From 1929 to 1976, the average workweek of fixed capital in manufacturing increased about 25 percent or at an average annual rate of 0.47 percent (Table 1). This rise reflects an increase in the number of hours worked by plants per day as a result of increased shift work partly offset by a decrease in the number of days worked per week. The overall rise occurred in the face of a decline in the average workweek of labor from a customary 50 hours per week in 1929 (actual hours were lower) to a 40-hour standard in 1976. The overall rise occurred even though there was little or no change in the length of the plant workweek for part of the capital stock that has typically worked around the clock.

From 1929 to 1976 the gross stock of fixed capital in manufacturing (Commerce) in 1972 prices rose by 166 percent or at an average annual rate of 2.10 percent. The 0.47 percent annual rise in average weekly plant hours over this period was thus 22 percent of the gross stock increase. In most 2-digit industries, increases in plant hours reinforced increases in capital stock.

The rise in average weekly plant hours provides a partial explanation for the drop of almost 45 percent in the ratio of fixed capital to output in manufacturing from 1929 to 1976. The drop is about 30 percent after adjusting for the rise in plant hours (Table 2).

TABLE 1—AVERAGE WEEKLY PLANT HOURS AND THEIR CHANGE, 1929–76, BY MAJOR INDUSTRY

	1929	1976	Percent Change
Food	88.0	90.3	2.6
Tobacco	49.2	104.4	112.2
Textiles	66.8	115.6	73.0
Apparel	46.0	46.3	0.6
Lumber	58.2	62.6	7.6
Furniture	50.3	52.6	4.6
Paper	128.7	139.6	8.5
Printing and Publishing	62.7	82.8	31.6
Chemicals	108.1	138.2	27.8
Petroleum	157.8	162.9	3.2
Rubber	103.7	120.0	15.7
Leather	49.4	45.6	−7.7
Stone, Clay, and Glass	104.6	119.3	14.1
Primary Metals	125.5	142.4	13.5
Fabricated Metals	55.6	77.8	39.9
Machinery	55.7	83.5	49.9
Electrical Machinery	49.6	77.0	55.2
Transportation Equipment	60.5	88.8	46.8
Instruments	59.5	76.6	28.7
Miscellaneous	49.3	62.1	26.0
All Manufacturing			24.7

Source: Estimates based on data from Bureau of Census and BLS.

Notes: 2-digit industries—1976, average weekly plant hours at 4-digit level weighted by book value of gross fixed assets (Census);—1929, average weekly plant hours at detailed industry level weighted by horsepower. For both years industry classifications have been modified. All manufacturing change: percent change in 2-digit averages as calculated above weighted by 1954 gross fixed assets in 1972 prices (BLS).

When changes in capital input are estimated by the change in the stock of capital in place, the effect of longer average weekly plant hours should be included in the change in total factor productivity. Using Kendrick's estimates of total factor productivity in 2-digit industries for 1929–48 and 1948–76, I regressed the 1929–76 change in plant hours on the change in total factor productivity. For the twenty 2-digit industries, $r=.45$, which is significant at the .05 level.

[1] The 1976 questionnaire instructs respondents to refer to the "duration the plant is open and operating" in reporting days per week and hours per day. Weekly plant hours were derived as the product of industry averages of these two items for "actual operations." In 1929, respondents were asked to report "normal numbers of hours plant was operated per day and per week." Shift data were also reported.

TABLE 2—1976 INDEXES OF MANUFACTURING OUTPUT, FIXED CAPITAL, AND CAPITAL–OUTPUT RATIOS, WITHOUT AND WITH ADJUSTMENT FOR CHANGES IN AVERAGE WEEKLY PLANT HOURS (1929 = 100)

	Fixed Capital			
	Without Hours Adjustment		With Hours Adjustment	
Output	Gross	Net	Gross	Net
480.8	265.8	254.7	331.6	317.6
	Fixed Capital-Output Ratios			
	55.3	53.0	69.0	66.0

Source: Output—Kendrick and Bureau of Economic Analysis; Capital Stocks—Bureau of Economic Analysis; Hours Adjustment—see All Manufacturing, Table 1.

The data on weekly hours of plant operations made possible an estimate of the growth of "continuousness" in manufacturing, which is an important aspect of modern technical change. A rise in the capital-labor ratio and a reduction in the time during which capital is idle are two of the most important results of the trend toward continuousness. Within each 4-digit industry in 1976 it was possible to obtain tabulations of plants working 150 hours or more per week and their share of employment. These detailed industry percentages were weighted by gross fixed assets to obtain 2-digit industry percentages and a percentage of all manufacturing. This yielded a figure of 28 percent for 1976. Similar tabulations for 1929 that made use of horsepower to combined detailed industries yielded a figure of 16 percent for that year. The difference between the figures would be smaller if the mix of industries at the 2-digit level were held constant.

To explain industry variations in the change in weekly hours of plant operations from 1929 to 1976, I made use of some broad facts that emerged from the statistical tabulations, some institutional developments and certain portions of the micro theory underlying shift work (see Robin Marris, Gordon Winston, and Roger Betancourt and Christopher Clague). Both cross-sectional and change versions were tried based on essentially the same group of industries, which numbered from 85 to 95. The cross sections, which were done separately for 1929 and 1976 and usually with the same variables, explained more than half of the variance in each year. But only the change version results are described here. The dependent variable is ordinarily the difference in average weekly plant hours in 1976 as compared to 1929. The independent variables, data for which came mostly from the census, are as follows:

1) The change in continuousness. Two versions of this variable were used. One, the share of employment in continuous plants, has already been described. A second was a dummy variable indicating the presence or absence of a single plant operating 150 hours or more per week in a given industry.

2) The change in capital intensity. The greater the capital intensity, the more likely the use of shift work. This variable is obviously related to continuousness. Because measures of capital intensity based on data that take no account of shifts can be misleading, I wound up using the ratio of kilowatt hours of all electricity—a proxy for capital services—to wage earner man-hours.

3) Changes in wage differentials for late shift work. At a point in time wage differentials for late shifts discourage shift work. We have two important facts about these differentials from the BLS (see Charles O'Connor). First, they are small relative to straight-time wages in this country. Second, shift differentials in manufacturing have not kept pace with wages generally, at least in the postwar period. Testing of the hypothesis was severely hampered by data limita-

TABLE 3—RESULTS OF EQUATIONS EXPLAINING THE CHANGE IN AVERAGE WEEKLY PLANT HOURS FROM 1929 TO 1976, WITH TWO VARIANTS FOR CONTINUOUSNESS VARIABLE

	Regression Coefficients using	
Constant	Dummy for (4)	Percentage for (4)
	−10.54 (1.2)	−10.83 (1.2)
(1) Capital intensity	.2434 (2.3)[a]	.1951 (1.4)
(2) Single-plant firms	−.2354 (2.0)[a]	−.2503 (2.0)[a]
(3) Women	−.3036 (1.8)[b]	−.3569 (2.1)[a]
(4) Continuousness	13.04 (3.0)[a]	.1716 (.9)
(5) Labor workweek in 1929	1.96 (2.6)[a]	2.2332 (2.9)[a]
R^2	.28	.21
N	88	88

[a] Significant at .05 level.
[b] Significant at .10 level.

tions and was confined to a very small subsample of the large sample.

4) Changes in relative importance of single-plant firms. A few industries like apparel and footwear operate their plants fewer hours per week today than in 1929. These are industries in which owners may provide a significant share of total labor input or may constitute the only managerial input available to the firm. In such firms the owners prefer leisure to the income that might be earned through longer hours on extra shifts. This influence would have diminished over time as the importance of small firms within industries diminished. This variable was measured by the change in the proportion of value-added accounted for by single-plant firms from 1929 to 1976.

5) Changes in the importance of women in the labor force. If women are the main source of added labor supply, but do not wish to work at night or are prohibited by law from doing so, industries may find themselves inhibited in their ability to use shift work. However, the force of this influence probably grew smaller over time as a result of the movement toward equality between the sexes.

6) The effect of the Wage-Hour Law. The Fair Labor Standards Act of 1938 probably hastened the adoption of shift work because it required that overtime be paid at a time-and-a-half rate. My hypothesis is that the industries that adapted most rapidly to the 40-hour week and multiple shifts, that is, to long *plant* hours, were those that started off with long *labor* hours and a large potential liability due to the new wage-hour legislation. I used the labor week in 1929 as a proxy for its length around the time of the passage of the Wage-Hour Law.

Table 3 presents the results of the change equations with two versions for continuousness. With the dummy version, which has the higher r^2, the coefficients on (1), (2), (4), and (5) have the correct signs and all are significant at the .05 level. The coefficient for the share of women in total employment is negative and significant at the .10 level. The comparatively low t-values for capital intensity and continuousness when continuousness is measured as a percentage are attributable to multicollinearity.

The coefficient on weekly labor hours in 1929 indicates that, for every hour by which the workweek of labor in 1929 exceeds 40 hours, the length of the plant workweek increased by 2.0 to 2.2 hours from 1929 to 1976.

Because of data gaps my analysis of the effect of late-shift wage differentials was limited to nineteen industries and was confined to 1976. Since the shift differential is positively related to the supply of labor for late shifts and negatively related to the demand for late-shift labor, I used a model with both supply and demand equations. However, significant coefficients for the differential variable were not obtained.

To summarize: the rise in average weekly plant hours in manufacturing from 1929 to 1976 is explained in part by the rise in capital intensity and a related increase in continuous operations in industry. Industries where single-plant firms predominate today are those where plant hours are short and probably reflect a preference by owners of small firms for leisure as against income. The decline in the importance of such firms has contributed to the lengthening in plant hours. The overtime provisions of the Wage-Hour Law also contributed to the trend toward longer plant hours. The evidence is not strong but the growing presence of women in employment seems to have inhibited the use of shift work for the period as a whole. The long-term decline in shift differentials relative to wages has probably had some influence on the spread of shift work even though the severely limited sample with wage differential data did not yield statistically significant results.

I have not yet attempted to relate these findings to those of my 1963 study. Today's results provide part of the explanation for the rise in average weekly equipment hours in manufacturing I had found from 1929 to 1954, but they will not explain any rise in the efficiency of capital utilization on a given shift. Still, the increases reported here could help explain an *apparent* decline in average lives of equipment in use. They could also be one of the explanations of why the stock of manufacturing plants has risen so little in the past 25 or 50 years.

Work is underway on estimates of weekly plant hours for periods between 1929 and 1976 as well as on the rest of the private business sector. Until the picture is more complete, we ought to reserve judgment about economy-wide issues, such as the extent to which the long-run decline in the workweek of labor may have been offset by a rise in the workweek of fixed capital. That problem was addressed in a 1956 paper by Charles Schultze, who surmised that from 1890 to the early 1950's about two-thirds of the decline in the workweek of nonfarm labor was accompanied by a drop in hours worked by fixed capital.

It is interesting to speculate about the long-term relative decline in late shift wage differentials, and the concomitant rise in shift work. Does it mean that the supply curve of labor willing to work evenings and nights has shifted out? And is this because of improved transportation and plant amenities that have reduced the real cost to the worker of night work? It may be that decisions by business to locate new plants designed to operate two and three shifts have been importantly influenced by the availability of a labor supply that was willing to work late shifts at low shift differentials. This could be the case where plants were located away from large cities and represented opportunities for farmers either to quit farming or to moonlight and thus improve their income. If so, increases in shift work in rural areas at relatively low shift differentials could be an important nexus between capital investment and the movement of labor out of farming. This rationale is consistent with Denison, who has attached importance to the shift of farm labor to nonfarm employment in explaining the rise in total factor productivity from 1929 to recent years.

REFERENCES

M. Abramovitz, *Resource and Output Trends in the United States Since 1870*, Occas. Paper 42, Nat. Bur. Econ. Res., New York 1956.

R. T. Betancourt and C. K. Clague, "An Economic Analysis of Capital Utilization," *Southern Econ. J*, July 1975, *42*, 69–78.

Edward F. Denison, *The Sources of Economic Growth in the United States and the Alternatives Before Us*, Suppl. Paper No. 13, New York 1962.

S. Fabricant, "Economic Progress and Economic Change," *34th Annual Report of the National Bureau of Economic Research*, New York 1954.

Murray F. Foss, *Changes in the Workweek of Fixed Capital: U.S. Manufacturing, 1929 to 1976*, American Enterprise Institute, Washington, forthcoming.

———, "The Utilization of Capital Equipment," *Surv. Curr. Bus.*, June 1963, *43*, 8–16.

D. W. Jorgenson and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Studies*, July 1967, *34*, 249–83.

John W. Kendrick, *Productivity Trends in the United States*, Princeton 1961.

Simon Kuznets, *Capital in the American Economy: Its Formation and Financing*, Princeton 1961.

Robin Marris, *The Economics of Capital Utilization*, Cambridge Univ. Press: Cambridge 1964.

C. M. O'Connor, "Late-Shift Employment in Manufacturing Industries," *Mon. Labor Rev.*, Nov. 1970, *93*, 37–42.

C. L. Schultze, "Some Economic Consequences of Long-Term Changes in Working Hours," mimeo., 1956.

R. M. Solow, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, *39*, 312–20.

G. Winston, "The Theory of Capital Utilization and Idleness," *J. Econ. Lit.*, Dec. 1974, *12*, 1301–20.

U.S. Bureau of the Census, *Census of Manufactures*, Washington 1929, *1*, 330.

______, *Survey of Plant Capacity*, Washington 1976, MQ-C1 (76)-1.

"The Measurement of Productivity," *Surv. Curr. Bus.*, May 1972, No. 5, Part II, *52*.

Slack Capacity: Productive or Wasteful?

By Walter Y. Oi[*]

"Full employment" is an ill-defined concept which means, of course, that slack is also poorly defined. Some idleness is surely part of an optimal allocation of resources. Slack may, however, be wasteful when it is generated by market imperfections. Slack capacity in the *U.S.* economy has apparently been declining over time, and this trend can be explained by received economic theory.

I. The Phenomenon of Idleness

Resources are never fully employed. An employed worker is "idle" fully three-fourths of the time. Over half of all factories are closed at nights and on weekends. Labor utilization rates, especially unemployment, have been analyzed in the literature, but W. H. Hutt was the one who emphasized the fact that idleness characterized the utilization of all resources, human and nonhuman. Idleness can occur in an equilibrium state as a result of an efficient organization of production. Other forms of "wasteful" slack may, however, be due to market failures. The observations that motivated this paper are described by four stylized facts.

A: Capital utilization rates have increased over the last four to five decades. Murray Foss (1963) reported that manufacturing equipment utilization rates had increased by one-third to one-half over the 1929–55 period. The corresponding increments were 20 percent in mining and 60 percent in electric power generation. In a recent study, Foss (1980) found that the workweek of fixed manufacturing plants (measured by hours per week) increased by 24.7 percent from 1929 to 1976. Similar upward trends were also reported by Charles O'Connor, and Roger Betancourt and Christopher Clague (1978).

B: The relative pay premium for late shift work has evidently declined. O'Connor, using BLS data, found that the average hourly earnings for the day shift rose by 31 percent over the 1960–67 period, while the late and night shift differentials increased by 14 and 15 percent, thereby reducing the size of the relative shift differential in spite of a relative increase in late shift employment.

C: The prices of capital goods measured in wage units have declined. Reliable statistics properly adjusted for changes in durability and quality are difficult to assemble, but the available data indicate a sharp fall in the relative price of new capital goods.[1]

D: Over time, capital goods are becoming less durable. According to Martin Feldstein and Michael Rothschild (p. 409), the average expected life of new nonfarm investment fell from 19.8 years in 1929 to 15.3 years in 1963. The age distributions for the stocks of automobiles, trucks, and aircraft which reflect actual scrapping (as opposed to expected lives), have exhibited leftward shifts over time.

These stylized facts can, I believe, be explained by received economic theory.

II. Optimal Utilization in a Putty-Clay Model

There is a class of deterministic models in which a higher utilization rate increases output, but it also raises some input prices. An optimal utilization rate balances these two opposing forces. Models differ in the

[*]University of Rochester.

[1] Data from the *Historical Statistics of the United States* reveal that the ratio of the wholesale value of a new car to the average hourly earnings of a production worker was $(V/W)=967$ hours in 1940. By 1970, the labor requirement for a higher quality new car was only 659 hours. The BLS producer price indexes for new capital equipment and for new construction rose more slowly than almost any series for wage rates.

specification of the production function and the way in which increasing use affects input prices.

In the Winston-McCoy model, a firm chooses instantaneous input rates for labor L and the capital-labor ratio, $Y=K/L$, which determines potential daily output, $\bar{Q}=Lf(Y)$. (Gordon Winston and Thomas McCoy, hereafter W-M, assume first-degree homogeneity.) *Ex ante* substitution is assumed, but once Y is chosen, the firm is frozen into a putty-clay technology. Actual output is proportional to the utilization rate, $Q=U\bar{Q}=ULf(Y)$. The daily rental of a machine, $R=V(r+d)$, is independent of U, but the daily wage is an increasing function of $U, W=W(U)$ with $W', W''>0$.[2] An efficient plan is one in which the capital-labor ratio Y and utilization rate U are chosen to minimize the total long-run cost for a given rate of output; i.e., minimize $C=(W+RY)L$, where $L=Q/Uf(Y)$. The first-order conditions are

$$(1a)\quad cUf'_Y=R,$$

$$\left[c=\frac{W+RY}{Uf(y)}=\text{minimum unit cost}\right]$$

$$(1b)\quad W'(U)=\frac{W(U)}{U}=\frac{RY}{U}$$

In equilibrium, the marginal value product of capital, cUf'_Y, is equated to the rental R, and the incremental labor cost for a longer workday is balanced against a falling "price" for the service flow from capital. The optimal values of Y^* and U^* are thus determined by relative input prices, $[R/W(U)]$ and properties of the production and wage equations. The W-M model yields two propositions, namely if the elasticity of substitution is less than unity, $\sigma<1$, the optimal utilization rate U^* will rise when (a) $[R/W(U)]$ is increased and (b) the amplitude of $W(U)$ is dampened meaning a smaller shift differential.[3]

In the W-M model, some slack is optimal because workers demand compensating wage differences for longer workdays. The model must, however, appeal to exogenous changes in the relative rental price of capital or in the premium required for longer shifts if it is to explain the secular trend in utilization rates.

The utilization rate can be varied by changing the length of a shift or the number of shifts. Models dealing with the latter discrete choice still assume a putty-clay technology but they replace the continuous wage equation with a shift differential; that is, the wage on the late shift is some multiple $(1+\alpha)$ of the day wage. The results are, however, qualitatively similar to the W-M model. Thus, Betancourt and Clague (1975) find that the conditions favoring multiple shift operation include (a) larger returns to scale, (b) a larger capital share of total costs, and (c) a small shift differential α. The discrete choice model is useful for econometric research, but it does not explain differences across shifts in the same factory; why is labor productivity lower on the late shift?

III. Variable Proportions and the User Cost of Capital

In several models, the capital utilization rate U appears as an argument of the production function, $R=f(U,K,L)$. The wage rate is constant, but R is an increasing function of U because greater use presumably increases the depreciation rate δ; i.e., $R=V(r+\delta)$, with $R_U=V\delta_U>0$. Unambiguous results can only be obtained by imposing restrictions on the production function. In the Epple and Zelenitz model, U and K do not appear as seperate arguments; only their product, $X=UK$, representing machine

[2] The daily rental R is determined by the price of the asset V, the interest rate r, and the constant depreciation rate δ. The wage equation could be derived from utility maximization where W' and W'' reflect a diminishing marginal rate of substitution of leisure for income. This is surely implied in the W-M model.

[3] An increase in R will reduce Y^*, but RY^* depends on σ. If $\sigma<1$, RY^* will rise, and U^* must be increased to restore equilibrium. In the special case of a Cobb-Douglas production function where θ is capital's constant share of output, equation (1b) can be simplified to $(1-\theta)W'(U)=W(U)/U$. Hence, U^* is independent of R for this production function.

hours, is entered in the production function, $Q=f(X,L)$. The minimization of total cost, $C=WL=RK=WL=R(X/U)$, yields the following first-order conditions:

$$(2)\quad Q=f(X,L),\ \frac{f_X}{f_L}=\frac{(R/U)}{W},\ \delta_U=\frac{r+\delta}{U}$$

Only the last equation determines the optimum utilization rate U^* which equates the marginal to the average user cost of capital. Thus, U^* is independent of W, V, and the elasticity of substitution σ; it is chosen to minimize the cost of the service flow (R/U).[4] A higher rate r will increase both U^* and (R/U^*) which leads to a lower capital service flow per worker,(X/L). These results differ from the W-M model, but the approach is qualitatively similar.

IV. Maintenance and Non-Linear Depreciation

In most models, the depreciation rate δ is assumed to be constant over the machine's life, but δ can depend on other variables. The assumed constancy of δ must be questioned in the light of data on the prices of used assets. The price of a used machine declines with increasing age in a non-linear fashion because of deterioration (due either to output or input decay) and technical obsolescence. The rental of capital must cover the amortization of the asset's value as well as maintenance costs μM that are required to "produce" a service flow from capital; i.e., $R=V(r+\delta)+\mu m$, where μ is the price for a unit of maintenance input m. The technology usually involves input decay meaning that older machines require more maintenance to sustain their productivity.

An optimal scrapping age (defining the economic life of a machine) is determined by the age profiles of used asset prices and maintenance costs. The differential between the value of a used machine V and its scrap value S decreases with increasing age, while maintenance costs rise. Following Richard Parks, a machine will be scrapped and replaced when maintenance costs exceed the net value of a machine, $\mu m \geqslant (V-S)$. Depreciation is endogenous. A fall in the price of new machines shifts the $(V-S)$ profile downward reducing the optimal scrapping age. The firm substitutes more of the cheaper capital for less maintenance labor. Taxes and subsidies could result in socially wasteful replacement policies.[5]

The forces that determine the optimum scrapping age are also responsible for the inverse relation between a machine's age and its utilization rate. This result follows from equation (1b) of the W-M model, where a newer machine will be more intensively utilized because it entails a higher capital charge, RY. If the user cost of capital is incorporated into the W-M model, this inverse relation is reinforced. If δ and m are increasing functions of U, the firm faces a rising cost of capital.

$$(4)\qquad R'_U=\frac{dR}{dU}=V\delta_U+\mu m_U>0$$

Equation (1b) is now replaced by

$$(5)\qquad W'(U)-\frac{W(U)}{U}+YR'_U=\frac{RY}{U}$$

Technical obsolescence accounts for a larger share of depreciation of newer machines implying that δ_U will be smaller. Further, the additional maintenance which accompanies more intensive use m_U is likely to be larger for older machines. Both factors lead to a more rapidly rising user cost R'_U for

[4] Dennis Epple and Allan Zelenitz were concerned with the way in which a rate of return constraint affects a firm's choice of the utilization rate U and built-in durability b. The rental rate of capital in their model is given by $R=V(b)[r+\delta(b,U)]$. They show that a regulated firm will *not* minimize R in maximizing constrained profits.

[5] When the federal capital grants subsidies paid for three-fourths of the price of a new bus, the optimal scrapping age which minimized the private bus line costs fell from 25 to 13 years. More frequent replacement was the means by which the company substituted cheaper capital for nonsubsidized maintenance labor. William Tye estimated that fully 25 percent of the capital grants subsidies were thus "wasted." Retirement of equipment prior to its socially optimal scrapping age is a form of "enforced idleness" that represents a deadweight welfare cost.

older machines. As a consequence, older machines are less intensively utilized.

V. Firm Size and the Organization of Production

Large and small firms in the "same" industry differ in several important respects; large firms (a) choose more capital-intensive production techniques, (b) utilize capital equipment and plant more intensively, (c) pay higher wages, (d) adopt less flexible production methods when judged by *ex post* substitution elasticities, and (e) produce standardized as opposed to customized products.

The optimum firm size defined by the output rate at which ATC is a minimum, reflects a balancing of scale economies from mass production and integration and of diseconomies arising out of the fixity of the entrepreneurial input. All firms might, for example, face the same production function, $Q=\phi(L,K,E)$ with small firms having small endowments of entrepreneurship $\bar{E}$. If $\sigma_{KE}>\sigma_{LE}$, the output expansion path in the range of diminishing returns will entail more capital-intensive techniques. This could account for a higher (K/L) ratio in larger firms.

Suppose that entrepreneurship is described by ability ε and managerial time T. An entrepreneur with greater ability, (large ε), faces a higher shadow price of time. He can economize on the use of this scarce input in at least two ways: (a) by substituting hired inputs for managerial time; or (b) by shifting the product mix away from goods and activity that place high demands on time. Monitoring input performance is a time intensive activity. A large firm with a higher shadow price for management time is thus more likely to adopt capital intensive methods because machines are more easily monitored than men. Production will be organized around teams and units whose performance can be more cheaply monitored. Further, monitoring outcomes is easier than monitoring inputs. The correspondence between outcomes and inputs can be more closely controlled by standardizing products and operating procedures. The latter usually involve rigid production schedules and fixed factor proportions which simplify the task of determining whether workers are following prescribed "efficient" production methods. Larger firms have to pay higher wages because workers must be paid compensating wage differentials to attract them into employments that require strict conformance with rigid, prescribed routines. With standardized operations, the large firm can design specialized equipment that reduces costs. If factor prices and product demands shift, the specialized capital entails a higher risk of technical obsolescence resulting in a higher capital charge RY. These higher capital costs prompt the large firms into choosing higher equilibrium utilization rates via multiple shifts.

Little firms are characterized by small endowments of entrepreneurial ability $\bar{\varepsilon}$ and low shadow prices for managerial time. They have a comparative advantage in time intensive activities like monitoring and observing individual worker performance. Jobs may be tailored to fit individual preferences. Workers can be reassigned and tasks redesigned to achieve better matches thereby enabling small employers to hire labor at lower wages. Because observation of individual performance and reassignment to new tasks are costly in terms of managerial time, larger firms devote more resources to recruiting and screening job applicants. We do indeed observe that hiring and recruiting costs in relation to total labor costs are higher for larger firms.

The product mix also varies with firm size. Customized goods are produced in small batches which are unsuited to capital-intensive, volume production. Small firms are better able to provide the flexible, adaptable organizations that can meet the shifting demands for customized, labor intensive goods. They will also buy general purpose, as opposed to specialized, equipment. Moreover, their lower labor costs give small firms an advantage in producing "maintenance." Hence, I would predict that smaller firms are more likely to buy used equipment and to scrap existing equipment at older ages. The user cost of capital R'_U, is thus likely to be higher for smaller firms leading to a lower capital utilization rate.

VI. Size of the Relative Shift Differential

The higher capital utilization rate has been achieved by increasing the proportion of workers on late and night shifts. Over the 1960–67 period, the pay differential for late shifts declined in spite of an increase in relative employment on late/night shifts. Marris argued that there may be some external scale economies which affect the private costs of working at night. When only a few work on late shifts, the size of the "night workers market" is too small to sustain frequent transit service and open shops at odd hours. When public transit serves as the travel mode, regular daytime work entails a lower "full cost" for a worktrip because of more frequent bus service. If, however, the worker travels by auto, congestion and parking costs are lower at night resulting in a negative full cost differential for late shift work. An expanded night workers market and the shift to increased auto use may have been responsible for the smaller compensating wage differential that is now paid to late shift workers.

VII. Concluding Remarks

Hutt identified eight categories of *idleness* of which five can properly be classified as wasteful.[6] Productive slack is exemplified by Hutt's concept of pseudo idleness:

> One of the most common forms of "pseudo idleness" is that which exists when resources are being retained in their specialized form (i.e., not being scrapped) because the productive service of carrying them through time is being performed. This condition exists when their capital value is greater than their net positive scrap value while their immediate hire value is nil.
>
> ...The bottling apparatus of a jam factory may be still for the early hours of each conventional working day. Such regular, recurring idleness can be confidently classified as "pseudo idleness." Spasmodic "pseudo idleness," on the other hand, can often be distinguished from idleness in other senses only with much uncertainty.
>
> [pp. 84, 85, 87]

Transaction costs and input supply prices during idle periods are simply too high to permit the emergence of positive "hire" or rental values.

The deterministic models examined in this paper generated equilibrium slack as part of an efficient organization of production. The equilibrium utilization rate depends on wage and user cost functions as well as on the age profiles of maintenance and depreciation. It seems reasonable to suppose that technical advances were responsible for the secular decline in the price of capital goods V, and that investments in human capital raised real wage rates. The higher wage increased the price of maintenance μ which in combination with a fall in V, led to earlier scrapping ages. The age distribution of the capital stock has thus shifted to the left meaning a less durable capital stock. Each machine yields a smaller service flow over its now shorter economic life, but the service flow per year off machine life is higher. Maintenance and non-linear depreciation can thus lead to a higher utilization rate in response to a decrease in (V/W).

In Section V, I sketched the outlines of a model in which the equilibrium size distribution of firms in a given industry is determined by relative factor prices, product demands, and the distribution of entrepreneurial abilities. A fall in the relative price of capital goods (V/W), redounds to the benefit of the larger, capital intensive firms. The relative price of standardized products will fall. Since large firms choose higher, equilibrium utilization rates, the mean utilization rate for all firms will rise. The equilibrium size distribution of firms will shift to the right, but this ought not to be interpreted as an anticompetitive change in the structure of an industry. The stylized

[6] Hutt's eight categories are valueless resources, pseudo idleness, preferred idleness which applies only to labor, participating idleness, enforced idleness, withheld capacity, strike idleness, and aggressive idleness. The last five categories are the results of monopolistic market conditions or government restrictions.

facts of Section I can thus be explained by received economic theory as the market responses to changing relative factor prices.

REFERENCES

R. R. Betancourt and C. K. Clague, "An Economic Analysis of Capital Utilization," *Southern Econ. J.*, July 1975, *42*, 69–78.

______ **and** ______, "An Econometric Analysis of Capital Utilization," *Int. Econ. Rev.* Feb. 1978, *19*, 211–27.

D. Epple and A. Zelenitz, "The Effects of Rate-of-Return Regulation on the Intensity of Use and Durability of Capital," *European Econ. Rev.*, 1979, *12*, 341–52.

M. S. Feldstein and M. Rothschild, "Towards an Economic Theory of Replacement Investment," *Econometrica*, July 1974, *42*, 393–434.

M. F. Foss, "The Utilization of Capital Equipment: Postwar Compared with Prewar," *Surv. Curr. Bus.*, June 1963, *43*, 8–16.

______, "Long-Run Changes in the Workweek of Fixed Capital," Amer. Enterprise Inst., xerox, Washington, July 1980.

W. H. Hutt, *The Theory of Idle Resources*, Liberty Press: Indianapolis 1977.

R. Marris, *The Economics of Capital Utilization*, Cambridge Univ. Press: Cambridge 1964.

C. M. O'Connor, "Late-Shift Employment in Manufacturing Industries," *Mon. Labor Rev.*, Nov. 1970, *93*, 37–42.

R. W. Parks, "Determinants of Scrapping Rates for Postwar Vintage Automobiles," *Econometrica*, July 1977, *45*, 1099–1115.

W. B. Tye, "The Economic Cost of the Urban Mass Transportation Capital Grants Program," unpublished doctoral dissertation, Harvard Univ. 1969.

G. C. Winston, "Capital Utilization in Economic Development," *Economic J.*, Mar. 1971, *81*, 36–60.

______ **and T. O. McCoy,** "Investment and the Optimal Idleness of Capital," *Rev. Econ. Stud.*, July 1974, *41*, 419–28.

U.S. Bureau of the Census, *Historial Statistics of the United States, Colonial Times to 1957*, Washington 1960.

Female Labor Supply in the Context of Inflation

By BETH T. NIEMI AND CYNTHIA B. LLOYD*

Inflation has become, in the 1970's, an important economic phenomenon that must be taken into account in analyzing the determinants of labor supply trends. Traditional labor supply theory has focused on the overriding importance of real wage growth as a primary determinant of both the long-run secular decline in hours of work and the postwar increase in the labor force participation rate of married women. Within this context, any labor supply effects of price level changes have been subsumed under the overall effect of real wages. However, despite unusually rapid price increases as well as stagnation in the growth of productivity and real wages in the last decade, the growth in women's labor force participation has continued and, if anything, accelerated in the last decade. Thus it is clear that, today, factors other than real wage growth must lie behind this continuing upward trend. Inflation is well worth exploring as a possible independent influence on labor supply, because of both its growing visibility and the frequency with which one hears comments along the lines of "in these inflationary times, a family needs two salaries to make ends meet."

Our objective in this paper is to explore the possible effects of inflation on women's labor supply trends. As a first step in this direction, we present some empirical results, relating changes in the labor force participation rates of women in various age groups to inflation as well as other independent variables, covering the period 1956–77. Inflation clearly appears to have an effect on labor force participation rates above and beyond the effect it generates through reducing the real wage. By next examining the primary sources of women's labor force growth in the last decade and their implications for long-run labor supply, we suggest the likely importance of inflationary expectations in sustaining the long-term growth of women's labor supply, particularly in the prime age group.

I. The Relationship between Inflation and Female Labor Force Participation

The inflation of the 1970's brought real wage growth to a halt and, in fact, wiped out the gains of the late 1960's. Measured in 1967 dollars, average weekly earnings in private nonagricultural industries were identical in 1979 and 1965. Real weekly earnings in 1979 were 2 percent lower than they were in 1970, and 7.5 percent lower than they were at their 1972–73 peak (see the *Economic Report of the President*, Table B-36, p. 245). With the growth in real wages thus reduced or eliminated, and market prices continuing to increase rapidly, the upward trend in female labor force participation might be expected to have slowed down as the relative advantage of market work has declined. The fact that this has not occurred suggests that inflation may have had some independent impact on labor supply that is not measured in the conventionally estimated labor supply relationship to the real wage, defined as the money wage divided by the Consumer Price Index (*CPI*).

There are three possible reasons why inflation might have an independent mea-

*Dr. Niemi died suddenly and unexpectedly, December 22, 1980. She was associate professor of economics, Rutgers University-Newark and vice president of Integral Research Inc., New York City. Dr. Lloyd is Population Affairs Officer for the Population Division of the United Nations. The views expressed are their own and not necessarily those of the United Nations.

sured effect, that is not reflected in the real wage, on labor supply. First, the *CPI* may be an inaccurate measure of the full increase in the cost of living or, alternatively, may overestimate this increase for some demographic groups and underestimate it for others. This is a problem caused by errors in the measurement of real wage changes and, although potentially significant, is not treated further in this discussion. The second type of situation occurs if people inaccurately perceive actual changes in real income. This phenomenon is referred to as "money illusion" if inflation is underestimated and thus real wage growth is overestimated, or "price illusion" if the growth in real wages is underestimated. This second case also amounts to a measurement or perception problem, one which we would expect to occur only in the short run. Finally, expectations concerning future wage and price growth may influence current behavior. If young people today aspire to increases in their standard of living over their prime working years comparable to those experienced by their parents during the mid-1940's to the mid-1960's (see Richard Easterlin), and if they expect present rates of growth in productivity and prices to continue into the future, the major remaining source of future increases in real family income in the long run would appear to be a shift from a one-earner to a two-earner family. The entry of wives into the labor force and increased weeks and hours of work on the part of part-year and part-time workers are the last few chips that many American families can cash in for a higher standard of living.

The recent effects of changes in real wages, employment opportunities and inflation on female labor force participation certainly merit examination. Some results of a recent study, in which we estimated labor supply elasticities for women with respect to these variables for the 1956–77 time period, using annual data, are relevant to the questions we are considering here. For a detailed description of the data, variable definitions, and estimation techniques used, see our 1981 article. We present here a summary of the observed effects of inflation on female labor force participation.

TABLE 1—THE NET EFFECT OF INFLATION ON FEMALE LABOR FORCE PARTICIPATION, 1956–77

	1956-77	1956-66	1967-77
All Women,16+	−.036	.932[c]	.073
	(−.421)	(2.39)	(1.10)
16-24	.394[b]	2.12[a]	.303[b]
	(2.32)	(4.29)	(2.46)
25-54	.180[c]	.439[b]	.219[b]
	(2.08)	(3.09)	(2.63)
55+	−.703[a]	.450	−.470[c]
	(−4.16)	(.548)	(−1.99)

Source: Our 1981 article, Tables 2 and 4.

Notes: The dependent variable in each equation is *lnLFP* for women in the age group in question. In addition to the *CPI*, the following independent variables were included in each regression: 1) the unemployment rate of prime aged men, 35-44; 2) an index of demand for women workers (i.e., an estimate of the relative employment demand for women workers based on the assumption that women's share of employment in each of the eleven major industries remains fixed, but the distribution of employment across industries changes over time); and 3) the income of year-round full-time women workers specific for each age group. All equations were run in *log*-linear form. The numbers in parentheses are the *t*-values.

[a]Significant at 1 percent confidence level.

[b]Significant at 5 percent confidence level.

[c]Significant at 10 percent confidence level.

The most interesting findings of this study were the positive and significant coefficients on the *CPI* for young and prime age women, which can be seen in the first column of Table 1. In fact, for women aged 25–54, the money wage and the *CPI* both affected labor supply positively. For older women, on the other hand, inflation had an independent negative effect on the labor supply. The hypothesis that money wages and the price level each affect labor supply only via their effect on the real wage, and thus the money wage and the *CPI* have equal and opposite effects on labor supply behavior, was consistently rejected. Overall, it appears that female labor force participation is increasing, in response to rising money wages and/or prices, more rapidly than would be predicted simply on the basis of changes in real wages.

Although the entire time period under consideration was characterized by steadily increasing female labor force participation, the 1970's also exhibited atypically high levels of both inflation and unemployment. The last two columns of Table 1 present the coefficients on the *CPI* for two eleven-year subperiods (1956–66 and 1967–77) of the twenty-two-year period in question. For young and prime age women, these coefficients are positive and significant in both subperiods, but smaller in the later years. The two periods differ sharply from one another with respect to the average quarterly percentage increase in the *CPI*: the average increase was approximately 0.5 in the first period and 2.3 in the second. It thus seems reasonable that expectation of a "normal" rate of inflation would be significantly different between the two periods, with larger absolute changes in the price level necessary to induce significant labor supply response in the second period relative to the first. For women of prime working age, although each 1 percent rise in the *CPI* induced a smaller increase in labor force participation in the second period than in the first, the total effect of inflation on labor force participation was considerably more pronounced in the second period, because the *CPI* rose much more rapidly.

II. Long-Run Labor Supply Implications

In attempting to understand the strong and continuing upward trend in female labor force participation—in particular whether it represents a fundamental shift in commitment to market work or merely a short-run response to a convergent set of economic inducements—as well as its possible links to the inflation of the 1970's, it is valuable to examine the composition of this increase and its relationship to other dimensions of female labor supply. Because the labor force participation rate weights equally people with different degrees of participation, it is impossible to tell without further examination what a change in labor force participation actually implies about total labor supply.

Although the relative contribution of different age groups to the growth in participation has varied considerably over time, every age group from 16 to 64 has exhibited a substantial increase over the past quarter century. Until 1965, the primary source of growth in the female labor force was concentrated in the age groups 35–64, and consisted largely of new entrants and re-entrants who had completed childbearing. Between 1965 and 1970, pronounced increases in the participation of women under 35 appeared for the first time, and since 1970, this age group has become the major source of continued growth. Between 1965 and 1979, the participation rate of women aged 25-34 rose from 38.5 percent to 63.8 percent (see *Employment and Training Report of the President*).

The effect of an increasing labor force participation rate on total labor supply will depend on what is happening to other dimensions of labor supply, and could be considerably weakened if these other dimensions are changing in the opposite direction. Trends in the continuity of labor force attachment as well as in average hours and weeks of work must be examined before we can speak with confidence about overall trends in female labor supply. Within the intermediate range (approximately 30 to 50 percent) over which female labor force participation rates have been rising during the past thirty years, one can easily visualize higher levels of participation linked to either increasing or decreasing labor force turnover. The question of whether today we have more women working, but for shorter and less continuous periods, or more women entering the labor force and then remaining there longer, is a crucial one.

Table 2 presents figures on labor force entry and exit for young, prime age, and older women. Net labor force growth is the difference between entries and withdrawals, and will occur more rapidly when the number of entrants increases, the number of exits decreases, or both. As can be seen in this table, entrants represent a decreasing fraction of the female labor force in both the young and prime age groups. Exit rates have also declined, indicating that more and more women are remaining in the labor force rather than dropping out. These data constitute but one example from the growing body of evidence that documents

TABLE 2—PROPORTION OF THE FEMALE LABOR FORCE MADE UP OF NEW ENTRANTS AND PROBABILITY OF LABOR FORCE EXIT BY AGE, 1967-79

	Entrants[a]		Exits[b]	
Year	Number (1,000's)	As Percent of Labor Force	Number (1,000's)	As Percent of Labor Force
16-24				
1968-69	3322	44.7	2799	37.6
1969-70	3135	39.7	2716	34.3
1970-71	3000	36.4	2752	34.4
1971-72	3106	36.0	2587	30.0
1972-73	3215	35.2	2708	29.6
1973-74	3211	33.4	2779	28.9
1974-75	2968	29.8	2682	26.9
1975-76	2917	28.4	2606	25.4
1976-77	3029	28.5	2624	24.7
1977-78	3152	28.5	2653	23.9
1978-79	2894	25.4	2706	23.7
25-59				
1968-69	3850	19.4	3121	15.7
1969-70	3657	17.9	3155	15.3
1970-71	3320	15.9	3047	14.6
1971-72	3462	16.2	2844	13.3
1972-73	3779	17.1	2994	13.6
1973-74	4053	17.7	3087	13.5
1974-75	3778	15.9	2931	12.3
1975-76	3870	15.6	2789	11.3
1976-77	4031	15.6	2902	11.2
1977-78	4325	15.9	2974	11.0
1978-79	4261	15.0	3030	10.7
60+				
1968-69	642	25.0	587	22.9
1969-70	687	26.1	619	23.5
1970-71	649	24.1	599	22.3
1971-72	681	24.9	632	23.1
1972-73	568	20.8	627	22.9
1973-74	546	20.5	629	23.6
1974-75	646	24.5	606	23.0
1975-76	588	22.0	565	21.1
1976-77	618	23.0	613	22.8
1977-78	693	25.4	617	22.6
1978-79	646	23.0	553	19.7

Source: *Employment and Earnings*, various dates.

[a] The number of entrants is calculated as the sum of the number who left and the net change in the civilian labor force.

[b] Estimated from persons out of the labor force who left a job within the previous 12 months. For example, to estimate the number who left between mid-1968 and mid-1969, an average was taken of the four quarterly figures for 1969. Because these four quarterly figures include persons who left jobs from the beginning of 1968 to the end of 1969, an average gives a more accurate estimate of the flow between the two midyears.

women's more continuous and longer term attachment to the labor force (see, for example, our book, ch. 2).

The final dimension of female labor supply to be considered is hours and weeks of work. There has been much discussion of the increasing importance of part time work opportunities for married women (see Nancy Barrett) and the growing number of part-time workers in the labor force. This can create the misleading impression that recent additions to the female labor force have

TABLE 3—PART-TIME STATUS OF THE CIVILIAN LABOR FORCE AND AVERAGE HOURS OF WORK FOR THOSE AT WORK, BY SEX, 1968-77

	Men: 16+		Women: 16+	
	Percent of Labor Force Part Time	Average Hours/Week	Percent of Labor Force Part Time	Average Hours/Week
1968	7.3	43.0	23.2	34.9
1969	7.7	42.9	23.5	34.9
1970	8.0	42.0	24.1	34.2
1971	8.1	42.2	24.4	34.3
1972	8.2	42.3	24.5	34.5
1973	8.0	42.4	24.6	34.4
1974	8.1	42.0	24.4	34.3
1975	8.2	41.6	24.2	34.1
1976	8.2	41.7	24.2	34.1
1977	8.4	41.9	24.2	34.2
1978	8.3	42.1	24.0	34.5
1979	8.1	42.0	23.8	34.5

Source: Unpublished BLS data.

consisted primarily of part-time workers. It is true that part-time work has become more prevalent among both men and women, but, as can be seen in Table 3, the increase in the relative importance of part-time work and the decline in average weekly hours do not appear to have been more pronounced for women than for men.

Evidence for the most recent two years even suggests that the growth in the importance of part-time work may now have leveled off for both women and men. It is also significant to note that women aged 25–34 were the one age-sex group for whom average hours of work rose, and the percentage working part time fell, over the 1968–77 period (see our 1979 book, p. 58). There is thus no evidence whatever that the vanishing dip in participation rates at ages 25–34 is being replaced by a dip in hours of work. Furthermore, although it is the case that on average women work four or five fewer weeks per year than men do, average annual weeks worked by women have increased slightly since 1960 (see our 1979 book, pp. 60–61). All in all, there is nothing to indicate that the impact of rising female participation rates on labor supply has been diluted by any countervailing trend in hours or weeks of work.

III. Conclusions

Whatever its multiple roots, the ongoing increase in female labor force participation in an inflationary context is clearly neither illusory nor temporary. Women's labor force participation and their full labor supply have been strongly correlated over the past twenty years; increased participation has not been offset by shorter hours or increased discontinuity. All the available evidence implies that recent increases in female labor force participation, particularly among women with young children, are the result of a trend toward long-term career commitment rather than an increase in marginal workers with high turnover rates.

While the recent stagnation in the real wage has not been the major source of discouragement for women that we might have predicted in the past, inflation is having an independent positive effect on female labor force participation. Since this phenomenon has persisted throughout the 1970's, the argument that it results from a misperception of the actual behavior of real wages (money or price illusion) is less than convincing. It appears that family labor supply behavior is influenced, not only by current wages and prices, but by the expectation that significant inflation will continue into the future, which may well be one factor contributing to the rise of the two-earner family.

REFERENCES

N. S. Barrett, "Women in the Job Market: Unemployment and Work Schedules," in Ralph E. Smith, ed., *The Subtle Revolution*, Washington 1979.

Richard Easterlin, *Population, Labor Force and Long Swings in Economic Growth: The American Experience*, New York 1968.

Cynthia B. Lloyd and Beth T. Niemi, *The Economics of Sex Differentials*, New York 1979.

Beth T. Niemi and Cynthia B. Lloyd, "Money Illusion or Price Illusion: The Effects of Inflation on Female Labor Force Participation," in Edward Marcus, ed., *Inflation Through the Ages: Economic, Social, Psychological and Historical Aspects*, New York 1981.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington 1980.

U.S. Department of Labor, *Employment and Training Report of the President*, Employment and Training Administration, Washington 1980.

A Times-Series Analysis of Women's Labor Force Participation

By JUNE A. O'NEILL*

The analysis of women's labor supply has stressed the importance of women's wage rates and husbands' incomes as factors influencing the labor force participation of married women. In this paper I address the question of whether the findings of cross-sectional studies are relevant for understanding the labor force behavior of women over time, particularly during the 1970's, a period of unusually slow growth in earnings and rapid growth in women's labor force rates. I also report on some estimates of my own, using aggregate time-series data. I conclude that the female wage rate and male income can explain surprisingly much of the trend over time, although the interaction of the process with divorce and other factors also influences the outcome.

I. Implications of Cross-Sectional Studies

Jacob Mincer's original framework for analyzing women's labor force participation made the point that the market wage not only influences the allocation of time between market work and leisure, but also between work in the market and work in the home. An increase in the market wage relative to the housewife's "wage" induces a substitution of market work for home work; the strength of the effect depends on the degree of substitutability between market goods and home goods. An increase in family income, other things the same, presumably has a positive effect on leisure, but may also indirectly affect the allocation of work time between home and market depending on the relative income elasticities of home-produced vs. market-produced goods. Thus, while an increase in income would be expected to result in a reduction of total work time, hours of work in the market may not decline much if the income elasticity for market goods is sufficiently stronger than that for home-produced goods.

Several studies have applied this conceptual framework to cross-sectional data, to estimate labor supply functions for married women. Mincer, in his 1962 study, uses a single equation model to analyze variation in married women's labor force participation in 57 large northern standard metropolitan areas (*SMSA*s) in 1950. His findings are in accord with a priori expectations: wives' wages have a strong positive effect on labor force participation while husbands' incomes have a negative, but weaker, effect. In addition, unemployment tends to discourage labor force participation. The effect of the wife's wage and husband's income alone were sufficient to explain half of the observed variation in labor supply.

Mincer's results were generally upheld in subsequent cross-sectional studies by Glen Cain and William G. Bowen and T. Aldrich Finegan, although the wage and income coefficients were weaker when the analysis was replicated with more recent (1960) data. Judith Fields, using 1970 data, found still weaker effects of wife's wage and husband's income. This result could reflect a real change over time in the labor supply function, if, as Fields suggests, women were significantly changing their work role orientation. Alternatively, it could reflect a change in the correlation matrix of the independent variables or other underlying statistical problems. Cain and Martin Dooley have attempted to improve the specification of the model by using a three-equation system in which wives' labor force participation, fertility, and wages are jointly determined, a formulation based on the presumption that these variables are endogenous. The Cain-Dooley results for 1970 do not differ sub-

*The Urban Institute. I am grateful to Rachel Eisenberg Braun for skillful research assistance.

stantially with respect to the labor supply function from those of Mincer.

A. *Application to Changes over Time*

To what extent can the labor supply functions estimated with cross-sectional data explain the change in women's labor force participation rates over time? In his early work Mincer found that changes in family income and the wife's wage could account for 70 percent or more of each decade's increase in labor force participation of married women, from the decade 1919–29 to the decade 1949–59. A similar exercise is conducted in Table 1 which compares the actual increase in married women's labor force rates with the predicted increase for each of three decades starting with 1947–57. Two predictions are given, one based on Mincer's equation, the other on the women's wage and male income coefficients from the Cain and Dooley study, ignoring the other variables in their model.

The labor force participation changes predicted by the two sets of variables correspond remarkably well to actual changes in labor force rates of married women. The Cain and Dooley coefficients produce larger predicted changes than the Mincer coefficients, accounting for 86 percent of the increase in the decade 1947–57 and 63 percent in the decade 1957–67, but overexplaining the change for 1967–77. Both sets of estimates correctly predict a smaller increase in labor force rates in the middle decade, 1957–67, than in the other two decades. This predicted slowdown in participation results from the relatively large absolute increase in male incomes compared to female earnings during the 1960's. The rate of growth in the economy slowed sharply in the most recent period, 1967–77. But, while the female wage rate increased more slowly than before, male income (including nonearned income) actually fell in real terms. As a result of this pattern, the negative effect of husbands' incomes reinforced the positive effect of wives' wage rates, producing an acceleration in women's labor force participation.

TABLE 1—ACTUAL AND PREDICTED CHANGES IN LABOR FORCE PARTICIPATION RATES OF MARRIED WOMEN, 1947–77 (Percentage Points)

		Predicted changes based on:	
	Actual Changes	Mincer $L=-.53H$ $+1.52W$	Cain and Dooley $L=-.52H$ $+1.6W$
1947-57	+9.6	+4.8	+8.3
1957-67	+7.2	+2.5	+4.5
1967-77	+9.8	+7.1	+11.4

Sources: Labor force participation data (U.S. Department of Labor); income and earnings (U.S. Bureau of the Census)

Note: L is the labor force participation rate of married women, H is the median income of all men, W is median earnings of women who worked year-round full time. Earnings and income are in 1949 dollars using Mincer's equations and 1969 dollars for the Cain and Dooley results.

One could not, of course, expect that estimates based on cross-sectional studies would perfecty predict changes over time. As Mincer noted, some factors which are fixed in cross sections may change over time, such as the relative prices of commercial substitutes for home-produced services and the rate of productivity change in the home relative to the market. A decline in the former would increase the relative attractiveness of market work at a given market wage, while an equivalent increase in home and market productivity would reduce it. However, technological change embodied in an increasing array of labor saving appliances for home production could induce a shift into market work, if the income elasticity of demand for home-produced goods is low. The situation would then be analogous to that in agriculture where rapid productivity advance combined with a low income elasticity of demand for farm products resulted in a decline in agricultural employment during the first half of the century.

Empirical evidence on home technology or the income elasticity of demand for home production is not readily available. There is evidence, however, that relative prices of capital equipment for the home have fallen.

The Consumer Price Index (*CPI*) for household appliances fell by 15 percent between 1957 and 1967, while the overall *CPI* increased by 19 percent; between 1967 and 1977, *CPI* increases for the two categories were 40 and 82 percent, respectively. At a given wage, such a fall in relative prices would provide an incentive to reallocate hours worked from home to market since it would induce the substitution of physical capital for labor in the home and increase the demand for wage goods.

Other changes over time may be stimulated by, and in turn stimulate, increases in women's labor force participation, a relationship which would not be present in cross sections. For example, the employment of women in the market may increase the risk of divorce; but an increased incidence of divorce also makes it risky to specialize in homemaking. In response to the increased supply of women workers, employers may adapt working conditions (such as hours) to women's needs, which would in turn encourage more women to enter the market.

Finally, the measures of earnings and income used in the cross-sectional analysis and in Table 1 are crude. Biasses may arise from the failure to adjust earnings for changes over time in skill or training and from the use of before-tax rather than after-tax income.

II. Findings from Time-Series

Although numerous labor supply functions for women have been estimated with cross-sectional data, I am not aware of studies using time-series. Time-series are often avoided because of a lack of observations as well as the belief that changes in attitudes or tastes cannot be measured and will obscure the effects of the economic variables. In this section I report on the results of regressing annual observations of women's labor force rates on women's wages, men's incomes, the unemployment rate of married men, the divorce rate, an industrial structure index and a time trend, for the period 1948 to 1978.

The model was estimated for all women and for married women, separately by age group: 20–24, 25–34, 35–44, and 45–54. Patterns of change in labor force participation over the period varied considerably among the age groups. Labor force participation rates for the oldest group increased by 13 percentage points in the decade 1948–58, but rose only slowly thereafter. Labor force rates for the two youngest groups, however, increased very little in the early decade but accelerated sharply in the most recent decade.

The rationale for selecting the two variables, women's wage rates and men's incomes, has been discussed earlier. The unemployment rate of married men is used as an approximation of purely cyclical disturbances. An industrial structure index is intended to measure changes in industrial composition affecting the demand for female labor, which has always been disproportionately employed in some industries and not others. A linear time trend is included so that variables such as wages and incomes that rise over time do not merely reflect a trend. The divorce rate is included as a proxy for the risk of losing the "job" of housewife. It is entered as a three-year average (ending in the year of the observation). The lagged form is used in an attempt to measure the effect of divorce on labor force participation, rather than the effect of participation on divorce. The number of children ever born per ten women is an additional variable included for the oldest group whose fertility is complete, and for the 35–44-year-old group whose fertility is almost complete. It is intended to be a measure of the cohort's labor force attachment in the past. Details of the data sources for the variables are given in the Appendix. All of the equations are linear, estimated by ordinary least squares.

Results for the period 1948–78 are given in Table 2 for all women and for the subset of married women. Although most women are married in the older age groups, there has been a trend towards delayed marriage among the youngest groups. In addition, divorces and separations have increased. To some extent the economic variables affecting labor force participation may operate through changes in marriage patterns, particularly at younger ages. The interest in the

TABLE 2—TIME-SERIES REGRESSIONS OF LABOR FORCE PARTICIPATION RATES OF ALL WOMEN AND MARRIED WOMEN, BY AGE, 1948–78

	Female Earnings[a]	Male Income[a]	Unemployment Rate	Divorce Rate[b]	Industrial Structure Index	Time Trend	Children Ever Born[c]	R^2 Corrected	*D. W.*
All Women:									
20-24	0.40	−1.2	−0.14	0.57	.37	.32		.988	1.6
	(3.5)	(−1.1)	(−0.6)	(3.1)	(0.35)	(1.8)			
25-34	0.31	−0.51	−0.26	0.57	−2.2	1.4		.990	1.6
	(2.6)	(−5.0)	(−1.1)	(3.2)	(−2.0)	(7.7)			
35-44	0.39	−0.31	—	0.38	−3.4	1.4		.992	1.4
	(2.5)	(−4.4)	(—)	(7.3)	(−4.8)	(11.4)			
	0.50	−0.12	−0.33	0.17	−0.15	0.87	−1.02	.996	1.8
	(4.5)	(−2.0)	(−2.7)	(3.0)	(−0.2)	(6.4)	(−5.2)		
45-54	0.49	−0.22	−0.28	0.14	0.84	0.75	−1.37	.990	1.5
	(2.5)	(−2.5)	(−1.3)	(0.7)	(0.9)	(4.1)	(−3.9)		
Married Women:									
20-24	0.55	0.01	0.33	0.64	−1.5	0.68		.978	2.0
	(2.5)	(0.5)	(0.7)	(1.8)	(−0.7)	(2.0)			
25-34	0.37	−0.43	−0.01	0.34	−2.41	1.48		.995	1.7
	(4.0)	(−5.3)	(—)	(2.4)	(−2.9)	(10.5)			
35-44	0.42	−0.29	0.05	0.25	−2.79	1.50		.994	1.6
	(2.4)	(−3.6)	(0.3)	(4.2)	(−3.5)	(10.8)			
	0.50	−0.15	−0.2	0.09	−0.42	1.1	−0.74	.995	1.8
	(3.1)	(−1.7)	(−1.1)	(1.2)	(−0.4)	(5.7)	(−2.6)		
45-54	0.58	−0.16	−0.65	0.42	3.2	0.43	−1.88	.993	2.4
	(2.5)	(−1.6)	(−2.6)	(1.7)	(2.9)	(2.0)	(−4.7)		

Sources: See the Appendix.
Note: *t*-values in parentheses.
[a] Shown in hundreds of 1964 dollars.
[b] Per 1,000 married women.
[c] Per ten women.

entire group of women stems from this factor.

Women's earnings have the expected positive effect and men's income the expected negative effect, and the results are in general statistically significant. The coefficient for own wage is somewhat lower and for husband's income somewhat higher for women 25 to 34 years, the age when small children are most likely to be present, thereby reducing the substitutability of market for home goods. As in cross-sectional studies, the unemployment rate usually has a negative effect but is seldom statistically significant. The divorce rate has the expected positive sign and usually has a stronger effect for all women than for married women. The effect for all women is expected to be stronger because an increase in the divorce rate increases the proportion of women who must support themselves, in addition to affecting the prospects of married women. The industrial structure index most often has a negative sign which is contrary to findings of cross-sectional studies. It may, however, reflect increases in the supply of women in other age groups. Past fertility behaves as expected and has a strong positive effect on the labor force rates of the older group.

The time variable is positive and significant. A number of omitted variables, however, may well be colinear with time such as years of school completed and changes in available work schedules, prices of household appliances, and attitudes toward work, as it becomes more commonplace for women to be working outside the home.

In conclusion, time-series results support the basic model which stresses the positive effect of the wage rate on the allocation of women's work time to the market and the negative effect of family income on work. The continued growth in women's labor force rates during the most recent decade despite a slowdown in the growth of women's real earnings can be explained by an even greater slowdown in husbands' total incomes combined with a sharp increase in marital instability. Comparing the different age groups, it is noteworthy that, over the period 1968 to 1978, younger women's labor force participation increased most rapidly while the real income of men under age 35 actually fell. During the same period the labor force rates of women 45 to 54 grew most slowly while men 45 to 54 experienced the greatest real increase in income of any age group. The fact that very high fertility cohorts were entering the 45–54 age group during this time reinforced the effect.

Appendix—Variable Descriptions and Data Sources

1): Labor Force Participation Rates: by age for all women and married women (U.S. Department of Labor).

2): Female Earnings (in 100's of $ 1964): for women 35–44 and 45–54, estimated median yearly earnings of year-round, full-time workers adjusted by average weekly hours of full-time workers (U.S. Department of Labor); and for women 20–24 and 25–34, women 25–34's hourly wage × 2000. Hourly wage data were calculated by William Butz of the Rand Corporation, who generously made them available.

3): Male Income (in 100's of $ 1964): Median yearly income of men by age (U.S. Bureau of the Census).

4): Unemployment Rate: of married men, spouse present (U.S. Council of Economic Advisers).

5): Divorce: moving 3-year average of divorce rates per 1000 married women 15 years and over (U.S. Bureau of the Census).

6): Employment Index: proportion of total employment by major industry (U.S. Council of Economic Advisers) weighted by percent female in industry in 1964 (U.S. Bureau of Labor Statistics).

7): Children ever born: cumulative birth rates per ten women, by age (U.S. Department of Health, Education and Welfare).

REFERENCES

William G. Bowen and T. Aldrich Finegan, *The Economics of Labor Force Participation,* Princeton 1969.

Glen C. Cain, *Married Women in the Labor Force: An Economic Analysis,* Chicago 1966.

______ **and M. D. Dooley,** "Estimation of a Model of Labor Supply, Fertility and Wages of Married Women," *J. Polit. Econ.,* Aug. 1976, *84,* S179–99.

J. Fields, "A Comparison of Intercity Difference in the Labor Force Participation Rates of Married Women in 1970 with 1940, 1950, and 1960," *J. Human Resources,* Fall 1976, *11,* 568–77.

J. Mincer, "Labor Force Participation of Married Women," in Gregg Lewis, ed., *Aspects of Labor Economics,* Universities-National Bureau Conference Series, No. 14, Arno Press: Princeton 1962.

U.S. Council of Economic Advisers, *Economic Report of the President,* Washington 1980.

U.S. Department of Commerce, Bureau of the Census, "Consumer Income: Money Income of Families and Persons in the United States," *Current Population Reports,* Series P-60.

______, *Trends in the Income of Families and Persons in the United States: 1947-1964,* Technical Paper No. 17, Washington 1967.

U.S. Department of Health, Education and Welfare, *Fertility Tables for Birth Cohorts by Color: United States, 1917-73,* National Center for Health Statistics, Washington 1976.

U.S. Department of Labor, *Employment and Earnings, United States, 1909-78,* Washington 1979.

______, *Employment and Training Report of the President,* Washington 1979.

______, *The Earnings Gap Between Women and Men,* Washington 1979.

The Economic Risks of Being a Housewife

By BARBARA R. BERGMANN*

To be a housewife is to be a member of a peculiar occupation—one with characteristics quite different from all others. The nature of the duties to be performed, the form of the pay, the methods of supervision, the tenure system, the "marketplace" in which "workers" find "jobs," and the physical hazards are all so different from conditions found in other occupations that one tends not to think of a housewife as belonging to an occupation in the usual sense. Yet being a housewife certainly meets the American Heritage Dictionary's definition of an occupation, as "an activity that serves as one's regular source of livelihood." In fact, to be a housewife is to be a member of the largest single occupation in the *U.S.* economy. Thus, it is certainly both legitimate and interesting to compare the advantages and disadvantages of the housewife's source of livelihood with those of other sources of livelihood.

Few economists have studied the economic aspects of being a housewife. Gary Becker's economium to the advantages of the division of labor among spouses addresses some of the issues, but its perspective seems to be that of a male member of a "traditional" family. More recently, Marianne Ferber and Bonnie Birnbaum, as well as Clair (Vickrey) Brown have made contributions in which the interests of each family member are recognized as distinct.[1] In this paper, I focus on the economic risks of the housewife occupation, which are shown to be very high relative to those of other occupations.

I. Characteristics of the Housewife Occupation

A. *The Duties*

The housewife's occupational duties—which we will define as the things she needs to do to keep her job—usually include cooking, dishwashing, housecleaning, laundry work, child care, and a "personal relations" component, which includes sexual relations. Both the sexual and nonsexual components of the duties are sources of economic risk. The nonsexual component of housewives' duties are broadly the same as the duties of paid domestic servants, although the housewife usually has more discretion than the servant, and a more responsible role with respect to the children and the finances. A housewife whose "job" ends, either at her own discretion or that of her husband, will probably have to enter some other occupation at least for a time. She will be faced with the fact that the alternative occupation most like the one she has left, and the one for which she has the most fitting recent experience, is one with both low pay and low status. Most housewives who lose their job or who quit do not become domestic servants. However, they will be at a disadvantage in the job market relative to their age group because of the failure in the eyes of employers to build up, during their service as housewives, that part of their human capital thought to be most serviceable on nondomestic jobs.

The sexual component of the housewife's duties also contribute importantly to the risks of the job. Like the airline stewardess, part of the housewife's job is being attractive. Unlike the stewardess, however, the housewife's duties clearly include cohabita-

*University of Maryland and The Urban Institute. Neither the University of Maryland nor The Urban Institute is in any way responsible for the contents of this paper. The views expressed here are solely my own. I would like to thank Deborah Freedman and Harriet Presser for comments on an earlier draft.

[1] Wassily Leontief once said that an input-output table for the American economy put together by a horse would look quite different from the one Leontief created. In "the new home economics," we are just starting to hear from the horses.

tion.[2] The woman who considers entering a housewife job, usually from a paid job, knows that cohabitation is a condition of keeping it, and considers the attractiveness of her suitor (or her husband, if she is already married) in deciding whether to accept the "job offer." Of course, sexual cohabitation in this context forms a perhaps vital part of the intimacy of the marriage relationship, with its presumption of caring, consideration, and long-run commitment. It is usually, at the outset at least, considered a highly valued fringe benefit rather than an onerous duty. However, the sex component of the housewife's duties, and the children who may appear as a result of it, make it difficult to go from one "job" to another within the occupation. The fact that the law makes marriages more costly and difficult to dissolve than are other employee-employer relationships also contributes to the difficulty of going from one housewife job to another.

The housewife's attractiveness to her husband can be thought of as a component of the human capital needed for her job, and she may be in the position of seeing this part of her portfolio of assets wane in value either gradually or suddenly. Her husband's attractiveness to her may also suddenly or gradually diminish, reducing the value of the intimacy fringe benefit, possibly changing it from positive to negative. These possibilities obviously make for high risk both with respect to "working conditions" and tenure.

Another component of a housewife's human capital which contributes to the value of her work is her identity as the mother of the husband's children, and thus as the person usually assumed to be most fitted to give them attentive and loving care. As the number of children born to marriages has on average diminished, and as the number of years in which a married couple has preschool children in the home has diminished, this component of the housewife's human capital disappears faster, leaving her more open to the threat of displacement from her "job."

As Ferber and Birnbaum have pointed out, the decline in the value of the housewife's services in the home occurs at a time when her husband's earnings and status are usually growing. The discrepancy in the economic position and in the social and sexual opportunities of a housewife and the man to whom she is married typically grows as they go through their forties.

B. *Physical Hazards*

The home is a risky place both for men and women. In 1977, 15.1 million women and 14.4 million men were injured in the home, while 2.4 million women and 9.0 million men were injured in paid jobs. Accidents are not the only source of injury; the chance of physical damage to women because of intentional human violence is far from negligible. A recent survey for the U.S. Department of Justice found that 4.1 percent of women living with a husband or male partner at the time of the survey had experienced severe physical abuse from him within the last twelve months and 8.7 percent had experienced it at some time. Severe physical abuse was defined as "being kicked, bit, or hit with a fist, being hit with an object, being beaten up, being threatened with a knife or a gun, or having a knife or a gun used against them." If in addition to the above, we include those women who had something thrown at them, or were pushed, grabbed, shoved or slapped, the proportion who experienced violence so defined comes to 10 percent for the previous twelve months, while 21 percent had experienced violence so defined at some time. Surveys which ask women about violent attacks on them by men living with them probably produce an underestimate of the extent of violence—they leave out women who have left their husbands, and attacks that women are ashamed to report.

In most occupations, physical assault on the job will be likely to result in criminal penalties for the perpetrator, but violence against housewives by their husbands results

[2] In all other occupations except prostitution the requirement that sex be part of the duties of the job has been defined by the courts as sexual harassment. See Catherine MacKinnon.

in few charges—the Justice survey found that only 9 percent of violent incidents are reported to police, and only 4 percent go to court. Police in most jurisdictions simply do not consider such attacks as criminal.

C. *The Pay*

The "pay for housewives" advocates, such as Carol Lopate, tend to ignore the fact that the housewife does receive a return for her work, perhaps because all or almost all of the pay takes the form of noncash benefits. Like the noncash benefits workers in other occupations get, they are untaxed.[3] The housewife's pay consists in room, board, a clothing allowance, medical care, all-expenses paid vacations, and the benefits she gets out of her own domestic services.

It is difficult to measure the housewife's pay directly, but it can be roughly assessed from published consumer expenditure data using a scheme in which a man's expenditure for his own clothing is taken to be an index of his standard of living. The cost of a housewife to a child-free married man can be taken to be the difference in family income on average between a married and single man with the same expenditures for male clothing. If a single man's clothing expenditure (C_s) is taken on average to be linearly related to his income (Y_s),

$$(1) \qquad C_s = a_s + b_s Y_s$$

and similarly a child-free married man's clothing expenditure (C_m) is on average

$$(2) \qquad C_m = a_m + b_m Y_m$$

then setting C_m equal to C_s we can express the cost of a nonworking wife to a husband as

$$(3) \quad W = Y_m - Y_s = \frac{a_s - a_m}{b_s} + \frac{b_s - b_m}{b_s} Y_m$$

The magnitudes of the parameters of (1) and (2) derived from the published aggregates in the 1973 *Consumer Expenditure Survey* give values for the constant term and coefficient of income in (3) remarkably close to zero and one-half, respectively. About half of a married man's income goes as "pay" to his wife.

We may carry the analysis still further by examining clothing expenditures for adult women. Again, the data indicate that single women dress about as well as child-free married women whose family income is about twice theirs. Since women who do full-time work for pay earn about 60 percent of what men do, a single woman who quits work to become a housewife appears to make on average a modest sacrifice in her standard of living, at least as measured by her standard of clothing consumption.

The noncash nature of the housewife's pay creates problems for her if conditions on her job are such that she wants to quit the marriage. It may be difficult or impossible for her to accumulate a cash reserve which would carry her through until she finds some other source of livelihood, usually a job in another occupation. If she can make such an accumulation, it may have to be done by stealth.[4] The "live-in" feature of the housewife's job increases the difficulty of quitting by increasing the size of the accumulation needed to change jobs. In most other occupations a person quitting a particular job does not have to move out of his or her present living quarters at the time of the quit—such a person can usually live for a while on the goodwill built up with the landlord and on the stocks of staples in the kitchen. It is conjectured that the reason a wife who is beaten by her husband may stay with him is that she has no place to go—and no resources to establish a place to live

[3] Rolande Cuvillier has for this reason characterized the housewife as an "unjustified financial burden on the community." She further suggests that agitation to provide the housewife with pay, disability insurance and better pension rights out of the public purse has more to do with insuring men that they will have domestic services than with improving the status of women.

[4] Some part of the usage of cash-back coupons offered on some products may be due to housewives' desire to have a source of cash not subject to their husbands' knowledge and supervision.

apart for herself and for any children she may intend to take with her. Even in less dire circumstances, the practical difficulty of setting up a new household may result in the imposition on and toleration by the housewife of circumstances few workers in other occupations are subject to.

II. The Economic Risks the Housewife Runs

If a risky activity is defined as one with a high variability in payoff, then the housewife's occupation is one of the riskiest. The variability of the housewife's pay is larger than the variability of her husband's, because it includes variation due to the possibility that the marriage will end, and that the pay she gets from him will cease. In this latter case, she may after a time be able to find a new job either as housewife or in some other occupation providing similar or even improved pay, but there will in most cases be a drop in economic status which is severe and prolonged.

Samuel Preston has estimated, based on disruption rates experienced in 1973, that 44 percent of marriages will end in divorce. The 1976 Census found the median interval between first marriage and divorce for women to be 7.3 years, which means that a substantial proportion of divorced women will have had relatively long marriages, and the housewives among them will have spent a substantial number of years out of the labor force. About 41 percent of the divorced women responding to the census had not remarried, and for those who had remarried, there was a median interval of 3 years between divorce and remarriage.

We may deduce from the 1976 Census data that on average a married woman runs a risk of divorce each year of above 2 percent. The risk of getting divorced in a year is far lower than the risk an employed person runs of suffering a spell of unemployment in a year, which was on the order of 14 percent in 1975. However, leaving or losing a job, difficult as that may be, is usually far less of a personal and financial trauma than ending a marriage is likely to be for a housewife. An important part of the financial trauma may relate to the expenses for children, since economic support from the husband for them will in a high proportion of cases disappear simultaneously with financial support for the wife. The 1979 Census reports that three-quarters of the mothers who were separated or divorced from their child's father received nothing from the father. Only 8 percent received $1,000 or more per child. Alimony is available with any regularity to an even smaller group.

The power that a husband now has to terminate his marriage to 'a housewife and thus to reduce considerably her standard of living and her status has effects on those housewives whose marriage has not terminated. First of all, there is the worry that the marriage may terminate. Second, the husband may use the implicit or explicit threat of leaving to achieve a dominance in the relationship, to the detriment of the housewife's feelings of well-being. Thus, the increasingly well-known risk of a bad outcome has the effect of reducing the value of a "good" outcome.

III. The Euthanasia of the Housewife?

If the housewife occupation has all of the disadvantages I have cataloged, why have so many people "chosen" it over other occupations? The answer, of course, is that all the people who "chose" it were women, and being women their alternatives were even worse, or were made to seem worse. In the past, the occupation of housewife had the character of a caste into which one was placed at birth. The socialization that female children received (and still receive) which makes membership in the housewife caste seem attractive and inevitable has been documented by Judith Long Laws. The ideology of romantic love and "Prince Charming" legend have played a part. For adult women, leaving the caste was discouraged by employment discrimination, by the social stigma attached to being a never-married or divorced woman, and by social pressures on wives not to seek paid employment.

In the era prior to the industrial revolution, most women worked on farms and contributed heavily to the output of goods in addition to providing housekeeping services. The housewife caste, in the sense of an occupational group devoting itself exclusively to domestic service, was greatly enlarged by the industrial revolution. As jobs off the farm were created, real wages for men got to a level such that many men were able to afford the services of a live-in domestic servant, who also served as a wife. This development segregated women's productive activities, but probably improved women's lives considerably. As technological change has proceeded further, the real value of cash wages available to women, although continuing to be far below mens' wages, have reached a level such that the alternatives to continued membership in the housewife caste have grown more attractive. Thus the same trends that caused the housewife occupation to grow are now causing it to shrink.

Although some women continue to enter the occupation of housewife directly from school and remain there for their lifetime, most women have only a spell as full-time housewives, starting at the birth of their first child. The number and lengths of women's spells in the housewife occupation are decreasing, and an increasing number of women are managing to get through a lifetime with no spell at all.

The decline in size of the housewife occupation will mean that some of the functions this occupation currently serves will disappear, and others will be served in other ways, through the purchase of market services and through the greater participation of husbands and children in providing domestic services. Isabel Sawhill has posed the question of whether the world would be a better or worse place if there were no full-time homemakers. Although economists do not usually ask such questions about products or occupations which changing technology or tastes have caused to decline, the question is in the minds of many, and is becoming a political issue. It is certainly true that, from a narrow point of view, a system in which the homemakers are women with impoverished alternatives serves the comfort and interest of men and male children, and the short-run comfort of female children. However, just as few would say that discrimination against blacks makes the world a better place, or even a better place for whites, so increasingly many are unwilling to say that a world of poor opportunities for women is a better place, or even a better place for men.

REFERENCES

G. Becker, "A Theory of Marriage: Part I," *J. Polit. Econ.*, July/Aug. 1973, *81*, 813–46.

C. (Vickrey) Brown, "Home Production for Use in a Market Economy," in B. Thorne, ed., *Rethinking the Family: Some Feminist Questions*, New York 1981.

R. Cuvillier, "The Housewife—An Unjustified Financial Burden on the Community," *J. Soc. Policy*, Jan. 1979, *8*, 1–26.

M. A. Ferber and B. G. Birnbaum, "The 'New Home Economics': Retrospects and Prospects," *J. Consumer Research*, June 1977, *4*, 19–28.

Judith Long Laws, *The Second X; Sex Role and Social Change*, New York 1979.

C. Lopate, "Pay for Housework," *J. Soc. Policy*, 1974, *5*, 27–31.

Catherine A. MacKinnon, *Sexual Harassment of Working Women; A Case of Sex Discrimination*, New Haven 1979.

S. H. Preston, "Estimating the Proportion of American Marriages that End in Divorce," *Sociological Methods and Research*, May 1975, *3*, 435–60.

Isabel V. Sawhill, "Homemakers: An Endangered Species?," *J. Home Econ.*, Nov. 1977, 18–20.

The American Heritage Dictionary of the English Language, Boston 1979.

U.S. Department of Justice, *A Survey of Spousal Violence Against Women in Kentucky*, Law Enforcement Assistance Administration, July 1979.

U.S. Bureau of the Census, "Divorce, Child Custody, and Child Support," *Current Population Reports*, P. 20, No. 84,

The Roles of Jurisdictional Competition and of Collective Choice Institutions in the Market for Local Public Goods

By DENNIS EPPLE AND ALLAN ZELENITZ*

Having discovered and neatly portrayed the conditions for efficient provision of public goods, Paul Samuelson concluded that there was no viable mechanism for eliciting the information about preferences required to determine optimal public goods supply. Much subsequent theorizing about public finance decisions may be viewed as a quest for some demand-revealing mechanism which would refute Samuelson's assertion. Charles Tiebout, in particular, countered with the claim that a viable mechanism did exist for determining the optimal supply of local public goods. In his view, self-interested individuals reveal their preferences for local public goods by their choice of jurisdiction. Tiebout reasoned that each individual "adopts," from the menu of local fiscal environments, that local community which most closely reflects his preferences. He further argued that the greater the number and variety of communities, the more efficient would the public goods provision be.

Without empirical support, this ingenious idea lay dormant for several years. The question remained as to whether the mechanism described by Tiebout, even if it existed, operated so that local jurisdictions effectively elicited and satisfied preferences for local public goods. Fundamental to answering this question was the necessity of designing an experiment which would reveal the result of "voting with one's feet." No such experiment existed until Wallace Oates (1969) reasoned that operation of the Tiebout mechanism would result in the capitalization of differentials in fiscal variables into property values. His test findings indicated that indeed such capitalization was evident. At first, the results obtained by others who repeated and extended his experiment were generally consistent with his findings and hence his interpretation of the link between the Tiebout mechanism and the capitalization of fiscal variables. It was not long, however, before an alternative interpretation was suggested by Matthew Edel and Elliot Sclar, who argued that in the long run the Tiebout mechanism would result in no capitalization of differentials in taxes and public service levels. Their interpretation was precisely the opposite of that offered by Oates, and their empirical work —which indicated little evidence of capitalization—supported their interpretation.

In retrospect, conflicting interpretations of the Tiebout hypothesis were almost inevitable given the absence of a formal model embodying the Tiebout mechanism. In an earlier paper with Michael Visscher, we argue that Tiebout's hypothesis is as follows: If individuals are able to choose from among a multiplicity of local jurisdictions, the result will be a Pareto-efficient provision of local public goods. We present two alternative models which differ in the technology for production of local public goods. One embodies the efficiency properties claimed by Tiebout, the other does not. Using these models we demonstrate that 1) Tax capitalization cannot be tested by the procedure proposed by Oates because the tax capitalization parameter is not econometrically identified—a possible reason for the contradictory empirical findings of Oates and of Edel and Sclar, and 2) A test of capitali-

*Associate professor, Carnegie-Mellon University, and assistant professor, Tulane University, respectively.

zation is feasible[1] but it is not a test of Pareto efficiency and hence not a test of the Tiebout hypothesis.

The above line of research is the outgrowth of what we believe to be the central premise of Tiebout, that Pareto efficiency is achieved because self-interested individuals reveal their preferences for local public goods by their choice of jurisdiction of residence.

An alternative approach in the quest for a demand-revealing process has been to attack this problem head-on, to focus upon the political institutions for collective decision making. Thus, jurisdictions may be regarded as behavioral entities which possibly weight the preferences of their citizens in selecting fiscal variables. The weighting process, most often conceived of as a voting mechanism, can then be judged with respect to the adequacy with which citizens' preferences are expressed in the actual fiscal decisions. The most widely known specification of governmental fiscal choice is the median voter model, which does connect the fiscal choices made by a jurisdiction's government and the desires of that jurisdiction's residents. William Niskanen's work on bureaucracy has stimulated interest in a broader investigation of governmental objectives, and formal models embodying a departure from the median voter framework are now being explored by Thomas Romer and Howard Rosenthal, Barbara Spencer, and others.

The above models of governmental institutions and mechanisms of collective choice do not, however, explicitly recognize an important characteristic of local governments, that each is in proximity to similar jurisdictions. Thus, these models are not strictly models of local government decision making. To the contrary, all of these models focus on governments and collective entities that are conceptually isolated decision-making units. Local governments play a prominent role in this line of research primarily because they are viewed as natural laboratories for testing theoretically derived hypotheses. It is at this interface with data for local governments that this second line of research touches on the line initiated by Tiebout. However, we view these alternative research approaches as deficient, since each tends to ignore the other.

Although not explicit about the mechanism of intrajurisdictional choice, Tiebout casts doubt upon the notion that "...government's revenue-expenditure pattern... 'adapt[s] to' consumers' preferences" (p. 417). Similarly, research on collective decision making takes scant note of the potential impact of interjurisdictional competition on intrajurisdictional choice.

The neoclassical theory of the firm provides an instructive reference. Individual firms pursue a self-interested profit-maximizing objective but competition among a large number of such firms results in efficient pricing and allocation. Does the analogy apply? Does competition among a large number of local governments assure efficient public goods provision even if governments of individual jurisdictions pursue objectives that are not necessarily public-regarding? If jurisdictional competition results in efficient provision of local public goods despite the pursuit of potentially perverse governmental objectives, can local jurisdictions provide a suitable environment for testing hypotheses about institutions of collective choice and governmental decision making?

Tiebout's assertion regarding optimality has been attacked on two distinct fronts. First, there are those (see, for example, Truman Bewley and the references therein) who question the existence of the technological conditions necessary for Pareto efficiency, even assuming that local governments are public-regarding. Typical of this line of attack are the suggestions that the diversity of local governments may be insufficient to meet the residents' heterogeneous diversity of demands; or that the tax and spending mechanisms may cause some commodities' (for example, housing's) price

[1] Epple (1980) argues that a test of capitalization of fiscal variables is a joint test of the following hypotheses: 1) Residents and potential residents are informed about fiscal variables in alternative jurisdictions, and 2) Government restrictions (for example, zoning) do not cause individuals to consume more housing services than they would in the absence of restrictions.

to differ for consumers and producers; or that the production function for local services depends on the local jurisdiction's particular population (compare Oates, 1981). Thus, local governments may be unable to attain Pareto efficiency as envisioned by Tiebout, notwithstanding each of these government's desire to attain that end. Second, governments may have goals which are not completely public-regarding, governmental agents may make choices in a self-serving fashion. Even if Pareto efficiency could be assured, the electorate's welfare may not be maximized. Much recent work on the collective decision-making mechanism has focused upon this second source of inefficiency, different goals of citizens and their governmental agents.

It is this second line which we investigate in our forthcoming paper. Our strategy is to specify a model with conditions that would appear to be ideal for the effective functioning of jurisdictional competition. We assume a metropolitan area inhabited by a population of identical individuals. The homogeneous land comprising the metropolitan area is divided into a parametrically determined number of equal-sized jurisdictions among which residents can move costlessly. Government services are produced with constant returns to scale. Since our goal is to study the workings of interjurisdictional competition, we assume that each jurisdiction has an entrenched government (i.e., a government not subject to electoral direction or control) that determines the tax rate on housing services and the level of government services in the jurisdiction. The government's power is limited to the choice of these two variables, and the government is required to raise taxes that are at least sufficient to pay government expenditures.

To cast the question as starkly as possible, we begin by assuming that each government seeks to maximize profits. That is, each government chooses the tax rate and level of services to maximize tax revenues less government expenditures. Each government is assumed to pursue a Nash strategy. We then let the parametrically determined number of jurisdictions approach infinity and examine the properties of the equilibrium for this limiting case. An analogous procedure applied to profit-maximizing firms would result in the zero profit competitive equilibrium. In the jurisdictional competition case, we find to the contrary that government profit does not go to zero as the number of jurisdictions approaches infinity. Though the result was, to us, unexpected, the intuitive explanation is quite simple. Even if its share of the total land in the metropolitan area is small, a government can expropriate a portion of the rents on the land within its boundaries; land is not mobile and cannot be relocated to escape taxation.

Though jurisdictional competition does not eliminate government profit, the total of such profit in an area is decreased by an increase in the number of jurisdictions. This result, too, is easily explained. For simplicity, consider the case in which all governments provide zero government services; all tax revenues are "profit." Given a downward-sloping housing demand function for the area as a whole and a limited number of jurisdictions, the demand function for each jurisdiction will also be downward sloping. This is because an increase in the gross-of-tax housing price in one jurisdiction will induce residents to move from that jurisdiction, thus causing the price in all other jurisdictions to rise. As the number of jurisdictions increases, the elasticity of housing demand in any one jurisdiction increases. An increase in the housing price in a small jurisdiction, by inducing some residents to move, will only cause a small rise in housing prices elsewhere in the metropolitan area because a small jurisdiction will house a small proportion of the metropolitan area's population. In the limiting case, any one jurisdiction encompasses an infinitesimal fraction of the land in the metropolitan area, and the demand for housing in the jurisdiction is perfectly elastic. In setting a property tax rate, the government of the jurisdiction is fixing the difference between the gross-of-tax and net-of-tax price of housing. A small jurisdiction faces a more elastic housing demand function than a large jurisdiction but the same housing supply elasticity as the large jurisdiction. Hence the government of a small jurisdiction has relatively limited ability to influence the gross-of-tax price and its choice of tax rate primarily affects the net-of-tax price. As a result, government profit pèr unit of land declines as the number of jurisdictions increases.

Our paper also explores a second closely related question. Can governments of different jurisdictions pursue different objectives, yet coexist in equilibrium? For example, can one government pursue profit maximization, another expenditure maximization, and still another property value maximization? We find that such governments can coexist in equilibrium even in the limiting case in which the number of jurisdictions approaches infinity. Again, the reason is that governments are able to expropriate a portion of the rents on property within their jurisdictions. The amounts they expropriate may differ with objective functions, but jurisdictional competition cannot eliminate such differentials. The population is mobile but the land is not; many important results flow from this crucial fact.

Our results show that the Tiebout hypothesis does not obviate the need to model intrajurisdictional governmental decision making. Our results also show that interjurisdictional competition can significantly affect the levels of taxes and services actually realized, whatever the governmental objective may be. Theoretical results based on a model of a jurisdiction considered in isolation may be significantly altered by considering the jurisdiction as one of many local governments residents may choose. If a model is to be tested using data for local jurisdictions in a metropolitan area, the interdependence among those jurisdictions cannot be ignored.

The above results have convinced us that what is needed is a merging of these two lines of research. Models of the mechanisms of intrajurisdictional choice must be integrated with models of interjurisdictional competition if a satisfactory theory of local public goods markets is to be obtained. To the extent that local governments provide the basis for empirical tests of theories of governments, such a merger is also important for the more general study of public goods supply by all levels of government. There are three avenues by which the integration of models of inter- and intrajurisdictional choice can proceed.

One strategy, employed in our paper with Visscher, is to assume the existence of equilibrium and to study the properties of the assumed equilibrium. Even where a proof of existence is absent, theorems of the following form can be instructive. If an equilibrium exists and assumptions X are satisfied, then the equilibrium must necessarily have properties Y. Conditions X characterize the environment and properties Y are the implications to be subjected to empirical investigation. This is less satisfying than the study of a model in which existence has been proved. However, where sufficient conditions for existence have not been established, the formal study of necessary conditions is much more fruitful than an entirely heuristic approach.

Most examples of failure of existence of an equilibrium among local governments assume a heterogeneous population and heterogeneous jurisdictions. Such problems can be avoided by employing a model in which agents and/or jurisdictions are identical. While equilibrium in such a model is not guaranteed to exist or to be unique, many existence problems can be avoided. Such an approach also has the advantage of generating a model with symmetry properties that can be exploited for comparative static analysis. This approach is employed in our 1980 paper and in Paul Courant and Daniel Rubinfeld. In the former paper the results were, if anything, enhanced by the use of homogeneity assumptions. If jurisdictional competition does not eliminate monopoly power when residents are homogeneous, it can hardly be expected to be more efficacious if residents are heterogeneous. Of course, there are a host of issues for which heterogeniety along some dimension is crucial, but we suspect that there are a large number of interesting issues regarding governmental behavior and jurisdictional competition that can be studied in a model in which strong homogeneity assumptions are employed.[2]

The third and most satisfactory approach is to explore the implications of models in which existence and possibly uniqueness has been proved. An important step in this direction has been made by Frank Westhoff. He proves existence of equilibrium in a model in which the median voter determines the tax rate and governmental service level within each jurisdiction and individuals are free to move among jurisdictions. Since, in his model, revenues are raised by a tax on endowments and there is no market for land, the model may be of limited use in deriving empirically testable propositions. The significance of the paper is that it sets forth a method of proof that may be applicable in models with greater empirical relevance. Epple, Filimon, and Romer, for instance, have succeeded in proving existence in a class of models in which a housing market exists and revenues are raised by a tax on housing service. Such models are instructive not only for their implications, but also because they clarify the requirements for existence. Hence, the issue of whether fiscal zoning or other *ad hoc* institutional factors are required to guaranty existence can be explored in a rigorous manner using such models.

To conclude, let us reiterate. A deeper understanding of resource allocation in the local public sector is dependent upon research approaches which merge the role of collective choice institutions with that of jurisdictional competition. Theoretical work is needed that explores the existence and properties of equilibrium in models that offer the potential for empirical testing. Such models must reflect the role of housing markets and the fixity of jurisdictional boundaries as well as the choice mechanisms within jurisdictions that determine taxation and the provision of local public goods.

[2] Some of these have been tentatively explored in our working paper.

REFERENCES

T. Bewley, "A Critique of Tiebout's Theory of Local Public Expenditures," disc. paper, Northwestern Univ., Feb. 1979.

P. N. Courant and D. L. Rubinfeld, "On the Measurement of Benefits in the Urban Context: Some General Equilibrium Issues," *J. Urban Econ.*, July 1978, *5*, 346–56.

M. Edel and E. Sclar, "Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 941–54.

D. Epple, "What Do Tests of Tax Capitalization Test?," working paper, Carnegie-Mellon Univ., Aug. 1980.

______**, R. Filimon, and T. Romer,** "Existence of Equilibrium Among Governments in a Metropolitan Area," working paper in preparation, Carnegie-Mellon Univ.

______ **and A. Zelenitz,** "The Implications of Competition Among Jurisdictions: Does Tiebout Need Politics?," *J. Polit. Econ.*, forthcoming.

______ **and** ______, "The Relation Between Welfare Maximizing and Property-Value Maximizing Governments: The Effects of Jurisdictional Competition," working paper, Tulane Univ., Mar. 1980.

______, ______, **and M. Visscher,** "A Search for Testable Implications of the Tiebout Hypothesis," *J. Polit Econ.*, June 1978, *86*, 405–26.

William A. Niskanen, *Bueaucracy and Representative Government*, Chicago 1971.

______, "Bureaucrats and Politicians," *J. Law Econ.*, Dec. 1975, *18*, 617–43

W. Oates, "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis," *J. Polit. Econ.*, Nov./Dec. 1969, *77*, 957–71.

______, "On Local Finance and the Tiebout Model," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 93–98.

T. Romer and H. Rosenthal, "Political Resource Allocation, Controlled Agendas, and the Status Quo," *Public Choice*, Summer 1978,*33*, 27–43.

B. Spencer, "Outside Information and the Degree of Monopoly Power of a Public Bureau," *Southern Econ. J.*, July 1980, *47*, 228–33.

C. Tiebout, "A Pure Theory Of Local Expenditures," *J. Polit. Econ.*, Oct. 1956, *64*, 416–24.

F. Westhoff, "Existence of Equilibria in Economies with a Local Public Good," *J. Econ. Theory*, Feb. 1977, *14*, 84–112.

On Local Finance and the Tiebout Model

By WALLACE E. OATES*

Much of the recent research on the local public sector takes as its point of departure the Tiebout model of local finance in which individual households seek out a community of residence that provides a fiscal bundle closely approximating their demands for local services. A central theme of this literature is the efficiency-enhancing properties of the Tiebout solution. In the "pure" case, it is a straightforward matter to show that a system, in which mobile consumers "shop" among a large group of local jurisdictions that offer a sufficiently diverse set of local public goods at a "tax-price" equal to marginal cost, will generate a Pareto-efficient outcome. The result is, in fact, a close analog to the private-market solution, for, as Charles Tiebout pointed out, "Spatial mobility provides the local public-goods counterpart to the private market's shopping trip.... Just as the consumer may be visualized as walking to a private market place to buy his goods, the prices of which are set, we place him in the position of walking to a community where the prices (taxes) of community services are set" (p. 422).

The pure model, however, involves a set of assumptions so patently unrealistic as to verge on the outrageous. In particular, Tiebout assumed a world of footloose consumers, who move costlessly among local jurisdictions in response *solely* to fiscal considerations; the Tiebout household is unconstrained by travel costs to a location of employment or by any other nonfiscal ties to a given locality. Moreover, access to each local jurisdiction in the system must be available at a tax-price equal to the cost of servicing the marginal consumer.

While the model thus generates some appealing sorts of results, the demands it makes on the nature of consumer behavior and institutional structure are formidable to say the least. The issue is whether or not the local sector in the real world is sufficiently "Tiebout-like" in its structure and operation to permit the use of the model for purposes of prediction *and* prescription.

This is a hard question. As the earlier quotation from Tiebout suggests, it is not too dissimilar from asking, "Is the private sector of the U.S. economy competitive?" As we know, there is not a simple answer to this query. For some analytical purposes (perhaps, for a broad view of the incidence of certain general forms of taxation) the answer may be yes; for others (as, for example, an antitrust investigation of a particular industry) the answer may well be no. Likewise, we are not likely to reach a definitive, general answer to the question of whether or not the local public sector is *sufficiently* Tiebout-like; the response will depend on the specific problem for which this query has relevance.

Nevertheless, there has been a considerable empirical (as well as theoretical) effort over the past fifteen years that explores the workings of the local sector from a Tiebout perspective. I wish in Section I to offer some brief observations on this work before proceeding in Section II to the issue of local production functions.

I

At the most basic level, we can simply look to see if, as a necessary condition for the operation of the Tiebout process, there exists enough diversity in the local sector to permit the kind of sorting out according to demands for local services that is envisioned in the model. Casual observation suggests an affirmative response, at least for most large metropolitan areas in the United States: a newly arriving household, for example, with a place of employment in the central city will typically have a wide range

*Bureau of Business and Economic Research, and department of economics, University of Maryland.

of suburban communities from which to select a residence. In a more systematic study of this issue, William Fischel presents a quite fragmented view of the typical suburban sector with a multitude of local jurisdictions exercizing both fiscal and zoning powers, and in which the concentration ratio (the land area encompassed by the four largest suburban governments) is relatively low. Enough competitors appear to exist for the process to work; however, as Dennis Epple and Allan Zelenitz (among others) have argued recently, sheer numbers, while constraining the choice of local officials, are not sufficient to ensure competitive outcomes in the local sector.

Another strand of empirical work involves a long series of capitalization studies that have examined the impact of local amenities and taxes on property values. Although the studies vary widely in choice and definition of variables (for example, output versus expenditure measures of amenities) and specific findings, the results on balance suggest strongly that fiscal differentials across neighboring jurisdictions tend to become capitalized into property values. The interpretation of this apparently straightforward result has turned out to be quite complex. On one issue, there is a consensus: capitalization of fiscal differentials is consistent with the view that consumers "shop" among local communities. People (not surprisingly) appear willing to pay more to live in jurisdictions that provide, in particular, such amenities as superior schools and greater safety from crime. The empirical literature thus provides some support for the operation of the "demand side" of the Tiebout model.

On the supply side, matters are less clear. Several authors (Bruce Hamilton, 1976b; Matthew Edel and Elliott Sclar; Epple, Zelenitz, and Michael Visscher) have pointed out that *full*-Tiebout equilibrium would imply an absence of any capitalization: with a perfectly elastic supply of local communities, the benefits from higher levels of amenities would be precisely offset by the associated increase in local tax bills. However, this "strong version" of the Tiebout model surely stretches reality. If we introduce some restrictions on the supply of communities (see Mark Pauly or Mahlon Straszheim) or certain forms of voting behavior on the part of residents (John Yinger), a "modified" Tiebout equilibrium will, in most instances, exhibit capitalization. Moreover, in these latter "weak versions" of the Tiebout model, the outcomes are no longer so clearly efficiency enhancing. In short, there appear no clear, unambiguous inferences to be made from the findings of capitalization of fiscal differentials as regards the efficient functioning of the local sector.

Exploring another implication of the Tiebout hypothesis in a recent and provocative study of several hundred towns in Pennsylvania, Howard Pack and Janet Pack have concluded that individuals within communities exhibit far too much variation in their demands for local services to be consistent with a Tiebout-like process. This is a tricky issue. First, Pack and Pack use income as a proxy for the (unobservable) demand for local services; income is, no doubt, positively correlated with demand but not perfectly so—even if demand were perfectly homogeneous in a town, we would expect to find a nonzero variation in household income. Second, and more fundamental, is the nature of the test. How much variation is too much to be consistent with a Tiebout world? The problem is that the null and alternative hypotheses are unclear. We might pose as the null hypothesis that the intracommunity variance in demand is at least as large as the variance in the metropolitan population as a whole; the Tiebout hypothesis would surely pass this test at a high level of confidence. This is admittedly a rather weak test, but at least one with a sound conceptual basis.

The efficiency properties of the Tiebout model also depend on marginal-cost pricing: the marginal resident must pay a fee equal to the cost of extending the local service to include his consumption. Tiebout assumed, in this regard, that the local service is subject to costs of congestion. Empirical demand studies (see, for example, Theodore Bergstrom and Robert Goodman) have tended to support this result: these studies find, in general, that they cannot reject the null hypothesis of a constant marginal cost for an additional consumer. While this is consistent with the Tiebout view, the mechanism of finance is more troublesome. Tiebout himself was not very explicit on this; he hardly mentions local taxation. However, most localities do not place a central reliance on user fees; they employ a variety of revenue sources, often relying heavily on property taxation. The introduction of property taxation links inextricably the issues of efficiency in local services *and* in housing markets. Hamilton (1975) has shown that the Tiebout model can be extended to a framework in which localities make use of property taxation *and* of a zoning ordinance that specifies a minimum level of housing consumption. While the Hamilton model generates a Pareto-efficient outcome, it makes even greater demands on reality than Tiebout: the Hamilton equilibrium entails communities that are homogeneous *both* in demands for local services and housing consumption.

Moreover, when we introduce further complications in terms of renters (who, many demand studies tell us, seem to believe that they pay lower taxes than owner-occupants), commercial-industrial property which assumes part of the local tax burden, and various intergovernmental grants, the precise link between the tax bill of the marginal consumer and the incremental cost of local services is broken. On this point, Pack and Pack cite a wide variation in housing values within communities, which they take as *prima facie* evidence that tax-prices vary substantially among residents; some of this variation may be offset through capitalization, but, if so, it comes at the expense of introducing inefficiencies into local housing markets (Hamilton, 1976a). In brief, it is unclear how closely the effective tax-price facing a potential resident reflects the marginal cost of local services.

II

While the Tiebout literature has at least addressed the issues examined in Section I, it has virtually ignored what I see as a central problem in local finance: the nature of the production function for local services. As noted earlier, Tiebout envisioned the provision of local services to be subject to costs of congestion; more specifically, he postulated a U-shaped cost curve with respect to community size, the low point of which served to define optimal community size. Most of the subsequent literature has simplified matters even further by taking the cost per person of a given level of local services to be constant with respect to community size; by assuming identical production functions across communities and with an appropriate selection of units, output in each jurisdiction becomes identical with expenditure per capita. While this procedure simplifies the analysis, it overlooks an issue with quite profound and troublesome implications for public policy.

My contention is that, for certain key local services such as education, public safety, and environmental quality, the production function contains as arguments not only the usual direct inputs of labor and capital, but also the characteristics of the individuals who comprise the community. For public safety, for example, a given input of police services will be associated with a higher degree of safety on the streets the less prone are the members of the community to engage in crime. Likewise, the more able and highly motivated are the pupils in a certain school, the higher will be levels of achievement.

Somewhat more formally (following David Bradford, R. A. Malt, and Oates), let I represent a vector of direct inputs into the production of local services. For schools, for example, this vector could have as elements the number (and quality) of teachers, schoolrooms, and books. The vector I maps into a vector D of "directly produced" services. For education, D might consist of providing a given number of students with a certain kind of instruction (for example, a specified series of "standard" mathematical lessons). In the case of public safety, we might associate these directly produced services with particular levels of surveillance.

However, what concerns the residents of the community is not the elements of D, but levels of final consumption: the quality of the schools in terms of student achievement, the degree of safety on the streets, and the physical attractiveness of neighborhoods. But these final outputs depend only in part on direct public inputs. For any given I vector, the quality of local schools will be better the more able are students; similarly, the level of public safety will be higher, *ceteris paribus*, the more law-abiding are residents.

In more formal terms, we can express the individual's utility function as $U=U(C_1,C_2,...,C_n,Z)$ where C_i is the level of consumption (final output) of the ith local service and Z is a composite private good. In turn, the production functions for the C_i are of the general form: $C_i=C_i(D_i,E)$ where D_i is the vector of directly produced services (a function of I) and E is a vector whose elements are the characteristics of the residents of the community.

My central concern here is with the role of the vector E in determining final outputs of local services. There is plenty of evidence of its importance. The Coleman report and subsequent empirical work attest to the overriding weight of the characteristics of pupils and their families in explaining levels of achievement in local schools. Likewise, population characteristics are typically the major explanatory variables in equations seeking to explain crime rates (see Oates). There is, I believe, little doubt over the moment of the elements of the E vector.

Moreover, this perspective on local production functions has two provocative policy implications. First, it points to an important role for local zoning ordinances as a means for regulating outputs of local services. The existing local-finance literature views the central function of zoning as basically that of excluding lower-income households that will not make a contribution to the local treasury commensurate with their share of budgetary costs. Exclusionary measures to this purpose constitute "fiscal zoning" (see Hamilton, 1975). The contention here is that local zoning regulations can also serve, if admittedly imperfectly, as a mechanism for controlling the composition of the local population so as to enhance the quality of local services; this is "public-goods zoning." Moreover, it may well be the case that, for services like education and public safety, the variables comprising the E vector dwarf in importance the budgetary inputs of the I vector. There may be only a comparatively limited capacity to improve the quality of the most important local services through the public budget. From this perspective, it is not hard to understand the jealousy with which local officials regard their zoning prerogatives. Zoning may be the one policy instrument they have to exert some control over the more important variables determining final outputs of local services. While this view may raise some thorny issues of social justice with difficult, and perhaps uncertain, normative implications, I would suggest at the same time that it does possess some positive explanatory power.

The second implication of this formulation of local production functions concerns the efficiency properties of the Tiebout model (see the appendix to Oates). In particular, matters become a good deal more complicated. Note that, in this framework, residents of a community are *both* consumers of and inputs into the local services in their jurisdiction. *In consequence efficiency in consumption and in production become inseparable problems*. The sorts of issues that arise in this context are perhaps best suggested by a provocative example. The importance of peer-group effects in schooling are well documented. However, in an intriguing econometric study drawing on an unusually rich body of data, Vernon Henderson, Peter Mieszkowski, and Yvon Sauvageau found for their sample that the peer-group effect (as measured by the mean *IQ* of the class in which a particular student is placed) is not only extremely important in determining achievement, but is non-linear: "The achievement of individual students rises with an improvement in the average quality of their classroom situations, but the increment in achievement decreases with the level of average class quality" (pp. 97–98).

The implication of this result is that a mixing of weak and strong students will improve the performance of the overall student population. This will, however, run counter to the interests of the more able students. Note also that it suggests an outcome that can easily be at variance with the sorting out of households according to demands for local services. There may, in this instance, be real tradeoffs both between efficiency in consumption and in production and also among the well-being of different individuals. More generally, the problem is that the efficient consumption of local services will typically require, along Tiebout lines, relatively homogeneous populations within each jurisdiction, while efficiency in production *may*, as in our example, point to considerably more heterogeneity.

The explicit recognition that the quality of local services depends on community composition as well as budgetary inputs admittedly complicates significantly the theory of local finance. However, the issues here have important implications both for the efficiency and equity aspects of public policy. In particular, I don't see how we can truly come to terms with such major concerns as the reform of school finance to provide equal educational opportunity from a perspective that focuses on variations in expenditure per pupil.

III

Both the literature surveyed in Section I and my discussion of local production functions in Section II suggest that the local public sector exhibits certain "imperfections" when measured against a standard of perfect economic efficiency. While this raises certain troublesome *and* intriguing issues concerning the actual workings of the local sector, we should be careful not to overreact to all this and effectively "throw out the baby with the bathwater." The Tiebout model does, I believe, generate some important *descriptive* insights; I have noted earlier the evidence supporting the operation of the demand side of the model—people appear to consider fiscal variables in their selection of a jurisdiction of residence. Moreover, in spite of the various imperfections of the system, the existence of choice among communities offering varying outputs of local services surely has some important efficiency-enhancing properties. Individual households not only have some discretion over their consumption of these services, but the competitive aspects of the provision of local services encourage a certain responsiveness to consumer tastes and put some pressure on local officials to seek out reasonably cost-effective techniques of production. While competition among local jurisdictions may not completely eliminate the potential for self-serving behavior among local officials, it surely does limit significantly the scope for such behavior (see Epple and Zelenitz).

REFERENCES

T. Bergstrom and R. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

D. Bradford, R. Malt, and W. Oates, "The Rising Cost of Local Public Services: Some Evidence and Reflections," *Nat. Tax J.*, June 1969, *22*, 185–202.

M. Edel and E. Sclar, "Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 941–54.

D. Epple and A. Zelenitz, "Competition Among Jurisdictions and the Monopoly Power of Government," disc. paper no. 79/1, Tulane Univ. 1979.

______, ______, and M. Visscher, "A Search for Testable Implications of the Tiebout Hypothesis," *J. Polit. Econ.*, June 1978, *86*, 405–25.

W. Fischel, "Is Local Government Structure in Large Urbanized Areas Monopolistic or Competitive?," unpublished paper, Mar. 1980.

B. Hamilton, "Zoning and Property Taxation in a System of Local Governments," *Urban Studies*, June 1975, *12*, 205–11.

______, (1976a) "Capitalization of Intrajurisdictional Differences in Local Tax Prices," *Amer. Econ., Rev.*, Dec. 1976, *66*, 743–53.

______, (1976b) "The Effects of Property Taxes and Local Public Spending on Property Values: A Theoretical Comment," *J. Polit. Econ.*, June 1976, *84*, 647–50.

Capitalization and the Median Voter

By JOHN YINGER[*]

The past decade has witnessed a strong interest among students of local public finance in models of voting with one's feet and in models of actual voting. Several authors, including Noel Edelson, Susan Rose-Ackerman, and Michael Lea, have recognized that these two types of voting must be considered simultaneously. The capitalization of local fiscal variables into house values, which is a by-product of voting with one's feet, influences the decisions of the median voter; and the pattern of local services that arises through actual voting influences the allocation of households to communities. A full-fleged merger of the two types of voting requires an analysis of capitalization in a model that considers both the housing market and the local voting process in a metropolitan area with diverse local governments financed by property taxes. Previous articles have focused on pieces of this puzzle. This paper reviews my attempt to bring these pieces together.

My analysis reveals that, regardless of supply responses, capitalization is a feature of long-run equilibrium. Furthermore, in the presence of capitalization, an efficient pattern of local services cannot be obtained through voting with one's feet, but must be obtained instead through actual voting. If preferences in a community are not too diverse, the median voter will pick the efficient level of services with or without capitalization. However, the median voter's choice may not be efficient in an extremely heterogeneous community. In addition, capitalization breaks the link between tax payments and the choice of a community, and thereby insures that the property tax is not a benefit tax; in other words, it insures that housing is underconsumed relative to other goods.

These results provide a framework for evaluating state and federal policies toward local governments. Without much heterogeneity within communities, one could achieve efficiency by eliminating the property tax or offsetting it with large subsidies to housing. These approaches are not realistic, however, and a second best solution is to cut back local services. The form of such a cutback is important. I show that tax limitations do not lead to efficiency, but that a second best solution could be obtained by redesigning intergovernmental grants.

I. The Capitalization Equation

In an urban area with diverse local governments, the price of housing services reflects households' valuations of local service-tax packages. To be specific, every house has a rental value equal to the market price per unit of housing services, which is a function of the level of local public services, multiplied by the number of units of housing services the house contains. The market value of the house is equal to the present value of the stream of rental values minus the present value of the costs of owning it, including property taxes. In symbols,

$$V=[P^*(E)H-tV]/r \tag{1}$$

or

$$V=P^*(E)H/(r+t) \tag{2}$$

where V is the market value of the house, P^* is the gross-of-tax rental value of a unit of housing services, H is the number of units of housing services, E is the level of local services per household, t is the effective property-tax rate, and r is the discount rate.

The value of a house can also be said to be equal to the present value of the stream

[*]Program in City and Regional Planning, John F. Kennedy School of Government, Harvard University.

of rental values *net* of taxes, or

$$V=P(E,t)H/r \tag{3}$$

where P is the net-of-tax price of a unit of housing services. From equations (2) and (3) it follows that

$$P(E,t)=P^*(E)r/(r+t) \tag{4}$$

The capitalization equations (2) and (3) are familiar because they apply to the valuation of any asset and form the basis for the income approach to property appraisal. Furthermore, they are equivalent to the housing bid functions derived from the utility maximization problem of a household that is searching for housing and can be generalized to consider many income-taste classes. (See my earlier paper.)

Matthew Edel and Eliot Sclar, and Bruce Hamilton argue that capitalization will disappear in the long run as housing suppliers respond to the price differentials that it implies. To be specific, developers have an incentive to build houses in communities with the service-tax packages preferred by households and to create, if possible, new communities with desirable service-tax packages. However, these supply responses are limited by the opportunity cost of converting land from nonresidential to residential use. Once all profitable conversion has taken place—that is, once suppliers are in long-run equilibrium—all remaining variation in service-tax packages will be capitalized into house values. No supply-side mechanism can eliminate this variation or produce housing until capitalization is competed away. (See my earlier paper.)

II. The Median Voter's Decision

Once a household buys a house and becomes a resident, it will vote for the service-tax package it would bid the most to obtain. With capitalization, Edelson points out, this objective is equivalent to maximizing property values. Voters do not bid for housing, but they do choose the service-tax package for which mover households will bid the most. Consider a homogeneous community in which local services are financed by a property tax, property is assessed at market value, and the cost of local services is $\phi(E)$ per household. In this case, the median voter, who is simply a representative voter, will try to maximize his or her property value, as defined by equation (2), subject to the jurisdiction's budget constraint:

$$\phi(E)=tV=tP^*(E)H/(r+t) \tag{5}$$

The first-order conditions of this problem and the derivative of (4) with respect to t imply that voters will pick the level of E at which

$$P^*_E H=P_E H(1+t/r)=\phi_E \tag{6}$$

where the subscripts indicate partial derivatives with respect to E. Note that $P^*_E H=P_E H(1+t/r)$ is the increase in the rental value of a house from another unit of E; that is, it is the marginal benefit to a household from local services. In addition, ϕ_E is the marginal cost of local services. Hence, equation (6) states that in a homogeneous community voters will choose the level of E at which the marginal benefit from local services is equal to the marginal cost.

This analysis can be extended to a heterogeneous community. Let $\hat{P}(E,t)$ be the bid function of the median voter. As long as residents' housing services are fixed, that is, with no additions or renovations, maximizing $\hat{P}$ is equivalent to maximizing the median voter's house value. In a community with N households, the median voter will therefore pick E and t so as to

$$\text{maximize} \quad \hat{P}(E,t) \tag{7}$$

$$\text{subject to} \quad \phi(E)=t\sum_{n=1}^{N}P^n(E,t)H^n/rN$$

The first-order conditions of (7) and the derivative of (4) with respect to t indicate that the median voter will pick the level of E at which

$$\hat{P}_E\overline{H}+(t/r)\sum_{n=1}^{N}P^n_E H^n/N=\phi_E \tag{8}$$

where $\overline{H}$ is the mean level of housing services in the community.

In long-run equilibrium, profit-maximizing housing suppliers will sell only to the highest bidder. Therefore, the market price of housing services will be the same for all households in a jurisdiction. Thus, the median house value divided by the mean house value, $\hat{V}/\overline{V}$, is equal to $\hat{H}/\overline{H}$ and we can rewrite (8) as

$$(9)\quad \hat{P}_E\hat{H}+(t/r)(\hat{H}/\overline{H})\sum_{n=1}^{N} P_E^n H^n/N = \phi_E(\hat{V}/\overline{V})$$

The expression $(\hat{V}/\overline{V})$ is the familiar tax-price for the median voter. This result indicates that the median voter is influenced not only by his or her own preferences and tax-price, but also by the distribution of preferences, that is, of $P_E^n H^n$.

An important special case, which is equivalent to the fixed-tax-share case analyzed by Edelson, arises when P_E is the same for all households in a jurisdiction. In this case, which involves heterogeneity only in housing services, (8) simplifies to

$$(10)\quad P_E\hat{H}(1+t/r)=P_E^*\hat{H}=\phi_E(\hat{V}/\overline{V})$$

The left-hand side of (10) is the median voter's marginal benefit from E; the right-hand side is his or her marginal cost. Because P is constant, (10) can be rewritten

$$(11)\quad P_E\overline{H}(1+t/r)=P_E^*\overline{H}=\phi_E$$

that is, the median voter sets the mean marginal benefit from services equal to the mean marginal cost.

III. Capitalization and Efficiency

In his famous article, Charles Tiebout not only provides insight into household moving behavior, but also argues that voting with one's feet leads to an efficient pattern of local services. Household shopping for a community, claims Tiebout, is analogous to shopping for a private good and therefore leads to efficiency. Local property taxes, which were not considered by Tiebout, complicate the efficiency debate because they effect both the level of local services and the consumption of housing. Wallace Oates argues that local property taxes do not alter the conclusion that household mobility leads to an efficient pattern of services, but that these taxes do lead households to consume less than optimal quantities of housing services. However, Hamilton shows that Tiebout's argument extends to a world with local property taxes. Each household picks the community where its valuation of the last unit of services is exactly equal to the property-tax increment needed to pay for it.

All of these arguments neglect capitalization, however. The authors either ignore capitalization or claim, incorrectly, that it will disappear in the long run. This neglect is important because capitalization shifts the grounds of the efficiency debate. Consider a class of households that live in several jurisdictions, only one of which has the optimal service-tax package for that class. Capitalization implies that the households receiving the less-than-optimal packages are compensated through a lower price of housing, and therefore have no incentive to move to the jurisdiction with the optimal package. In other words, households reveal their preferences through their bids for housing, not through their choice of a jurisdiction. Therefore, efficiency cannot be achieved through the economic process of voting with one's feet, but depends instead on the much less reliable political process of actual voting.

Recent articles by Edelson and by Jon Sonstelie and Paul Portney contain the argument that local voting will often lead to an efficient level of local services. To evaluate this argument, let us examine the case where voters have different levels of housing services but all place the same value on additional services. Consider the Samuelson efficiency condition for a collectively consumed good, such as E:

$$(12)\quad \sum_{n=1}^{N} MRS_{E,Z}^n = MRT_{E,Z}$$

where N is the number of households in the jurisdiction, Z is a composite consumption good, MRS denotes a marginal rate of substitution, and MRT denotes a marginal rate of transformation. If local governments produce E in a cost-minimizing manner, and if the market for Z is competitive, then producer behavior insures that $MRT_{E,Z}=N\phi_E/Q$, where $N\phi_E$ is the cost of producing another unit of E for each household and Q is the price of Z. Furthermore, (11) and the usual marginal condition for Z imply that each household sets $MRS^n_{E,Z}$, which is the ratio of the marginal benefits from E and Z, equal to ϕ_E/Q. Hence,

$$(13)\qquad \sum_{n=1}^{N} MRS^n_{E,Z}=N\phi_E/Q=MRT_{E,Z}$$

The level of E is efficient relative to a nonhousing composite good.

This result does not establish overall efficiency, however, because it does not consider the efficiency condition between E and housing. Under the above assumptions, producer behavior sets $MRT_{E,H}=N\phi_E/P$. Furthermore, with a property tax, households set the marginal benefit of housing equal to the price of housing plus the property tax, or $P(1+t/r)$. It follows that $MRS^n_{E,H}=\phi_E/[P(1+t/r)]$ and

$$(14)\qquad \sum_{n=1}^{N} MRS^n_{E,H}=N\phi_E/[P(1+t/r)] < N\phi_E/P=MRT_{E,H}$$

Thus, housing is underconsumed relative to local services. A similar argument shows that housing is underconsumed relative to the composite good.

In sum, a system of local governments financed by property taxes is not efficient because it involves underconsumption of housing. Remember that this result depends on capitalization. Without capitalization, Hamilton shows that the property tax is linked to the service level through household shopping for a jurisdiction. Capitalization breaks this link because it makes households indifferent across a set of jurisdictions and thereby "releases" the property tax to distort housing consumption. In effect, capitalization reinstates Oates' argument that the property tax distorts the housing market.

IV. Implications for Public Policy

The first best solution to the inefficiency in a system of local governments is to eliminate the property tax (and replace it with a nondistorting tax) or to offset the property tax through housing subsidies (paid for by a nondistorting tax). However, despite the fact that the property tax is unpopular and that subsidies flow to housing directly and through the income tax, it is unrealistic to expect these first best solutions to be complete.[1] Furthermore, in most states the composite good is subject to a sales tax. Hence, policy discussions need to take place in the realm of the second best. Because of existing institutional constraints, in other words, local services are overconsumed relative to other goods and policies designed to cut back local services are appropriate.

One must sail carefully through these second best waters. In the first place, my analysis applies only to owner-occupied housing. I have not derived results about voting in largely renter communities and therefore cannot make general prescriptions for service cutbacks.

Second, the attainment of efficiency requires that services be cut back in a particular way, namely so as to satisfy the local-services/housing efficiency condition. If θ is the proportion of property taxes not offset by subsidies, then this condition will be met if, in the jth jurisdiction, E is set so that

$$(15)\qquad P_{Ej}^{*}\overline{H}=\phi_{Ej}(1+\theta t_j/r_j)$$

Tax limitations clearly do not achieve this

[1] Because both property taxes and mortgage interest payments are deductible, income tax deductions may more than offset property taxes for some households. Consider a household in a 33 percent tax bracket with a fully mortgaged house at an interest rate five times the tax rate. In the early years of the mortgage, the subsidy is approximately $(.33)(tV+5tV)=2tV$. But interest payments decline over time and many homeowners do not itemize deductions, so that these deductions do not fully offset property taxes for many homeowners.

objective. With an arbitrary limit on t, and hence on E, one cannot expect equation (15) to be satisfied for any jurisdiction.

An alternative approach is to alter voters' incentives through state matching grants. For example, a state could, in effect, tax local governments for spending too much. With a matching rate of m_j, the marginal cost of services is $\phi_{Ej}/(m_j+1)$, so that the median voter will set $P_{Ej}^*H_j=\phi_{Ej}/(m_j+1)$. To reach the second best solution defined by (16), therefore, a state must set matching rates so that $\phi_{Ej}(1+\theta t_j/r_j)=\phi_{Ej}/(m_j+1)$, or

$$m_j=\left[r_j/(r_j+\theta t_j)\right]-1 \tag{16}$$

Thus, the state "tax" is implemented through a set of negative matching rates, which increase in absolute value as t_j increases. These negative grants would be added to existing grants, so that total grants need not be negative.

Third, one must account for heterogeneity within jurisdictions. As shown by Edelson, and as implied by equation (8), local voting will not necessarily lead to efficiency if a jurisdiction contains diverse preferences for services.[2] In fact, heterogeneity could lead some jurisdictions to underconsume local services. Equation (16) should be adjusted to account for this source of inefficiency.

Finally, this analysis does not consider equity. Service cuts obviously have important equity consequences, and any policy response to overspending on services should balance equity goals and efficiency.

Many economists appear to be puzzled by the recent rash of property-tax limitations because of the widespread conclusion that a system of diverse local governments will often lead to an efficient pattern of local services. Ironically, however, this efficiency may help to explain the tax revolt; local services satisfy the pure efficiency conditions, and are therefore overconsumed relative to housing, which is subject to the property tax, and to other goods, which are subject to the sales tax. In order to contribute to the policy debate surrounding tax limitations, economists need to take this second best situation seriously. We should design policies that recognize not only the distorting effects of property taxes, but also the inefficiency caused by intrajurisdictional heterogeneity.

[2] This result is well established in voting models without capitalization. See John Akin and Douglas YoungDay and the references cited therein. Under reasonable assumptions about the distribution of preferences, voting leads to underspending in large, heterogeneous jurisdictions such as central cities. It may be efficient, therefore, to cut services more in the suburbs than in the central cities.

REFERENCES

J. S. Akin and D. J. YoungDay, "The Efficiency of Local School Finance," *Rev. Econ. Statist.*, May 1976, *58*, 255–58.

M. Edel and E. Sclar, "Taxes, Spending and Property Values: Supply Adjustment in a Tiebout-Oates Model," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 941–54.

N. M. Edelson, "Voting Equilibria with Market-based Assessments," *J. Public Econ.*, Apr./May 1976, *5*, 269–84.

B. W. Hamilton, "Zoning and Property Taxation in a System of Local Governments," *Urban Studies*, June 1975, *12*, 205–11.

______, "The Effects of Property Taxes and Local Public Spending on Property Values: A Theoretical Comment," *J. Polit. Econ.*, June 1976, *84*, 647–50.

M. J. Lea, "Local Public Expenditure Determination: A Simultaneous Equations Approach," *Proceedings National Tax Association*, 1978, 131–36.

Wallace E. Oates, *Fiscal Federalism*, New York 1972.

S. Rose-Ackerman, "Market Models of Local Government: Exit, Voting and the Land Market," *J. Urban Econ.*, July 1979, *6*, 319–37.

J. C. Sonstelie and P. R. Portney, "Profit-Maximizing Communities and the Theory of Local Public Finance," *J. Urban Econ.*, Apr. 1978, *5*, 263–77.

C. Tiebout, "A Pure Theory of Local Expenditures," *J. Polit. Econ.*, Oct. 1956, *64*, 416–24.

J. Yinger, "Capitalization and the Theory of Local Public Finance," disc. paper, John F. Kennedy School of Government, Harvard Univ., Jan. 1980.

ISSUES IN GOVERNMENT ECONOMIC REGULATION OF TRANSPORTATION IN THE 1980'S

Antitrust Standards and Railway Freight Pricing: New Round in an Old Debate

By JOHN C. SPYCHALSKI*

Collective ratemaking has been practiced in various forms by American railway firms for more than a century. Its emergence antedated passage of the Sherman Act, and its survival despite both that legislation and subsequent federal and state antitrust laws stands out as a unique phenomenon in the annals of social control of business. Major challenges to collective railway ratemaking's continuance have occurred at three different periods. The most recent began with enactment of the Railroad Revitalization and Regulatory Reform Act of 1976 (4R Act), which narrowed the scope of antitrust immunity that had been accorded to rate bureau-based pricing by the Reed-Bulwinkle Act of 1948. Foremost among the 4R Act's changes in immunity were the prohibition of 1) collective determination of *particular* rates on single line movements, and 2) collective pricing of interline movements by railways which cannot "practicably participate" in such movements. Interpretation of the 4R Act's antitrust immunity provisions, to delimit their application, occupied a four-year proceeding in which the ICC, in a decision served August 13, 1980, issued a relatively narrow grant of immunity. The ICC also declared that the 4R Act prohibited discussion, as well as voting, on single line rates. In addition, it initiated a move toward elimination of antitrust immunity for collectively determined general rate increases and decreases, and broad changes in tariff conditions on single line traffic—*despite* a 4R Act provision that the prohibition on single line rate voting and the restriction on joint-line rate voting shall *not* apply to such general rate or broad tariff changes. The ICC, in essence, took the view that *intra*modal railway rate competition, governed largely by antitrust standards, is workable.

The ICC's decision imposes a standard of commercial behavior upon railways which, to the limited extent that it ever existed, was last seen in the nineteenth century. This marked departure from ingrained practice poses the question of whether it holds perceptible potential for improving efficiency in both transport and related sectors, as its proponents envisage. Scrutiny of conditions pertaining to this question, and to prospective consequences of the ICC's decision, occupies the remainder of the paper. Doing so requires the review of some very old but oft-slighted economic and institutional traits of rail transport.

I. Basic Cost and Market Characteristics of Rail Transport

The rail supply function is and always has been characterized by 1) high fixed costs vis-à-vis total costs within a period such as a fiscal year; 2) heavy investments in long-lived specialized assets; 3) marked rates of decrease in short-run average cost, but inconclusive evidence of long-run cost behavior; and 4) a large proportion of costs which cannot be traced causally to discrete sales units such as specific carload or trainload movements. The precise dimensions of such indivisible costs are difficult to ascertain. Suffice it to say that they will remain significant so long as the railway industry's traffic base includes numerous carload-sized consignments of a multitude of different types of commodities moving between thousands of different origin and destination points within a system which, by its very nature,

*Professor of business logistics, College of Business Administration, Pennsylvania State University.

occasions substantial common and joint costs.

At face, rail market structure—viewed *intra*modally in terms of the demand for carload or trainload (rather than *TOFC/COFC*) movements of particular commodities—appears to be a blend of 1) pure monopoly, when a particular shipper and consignee are linked by only one single line or joint-line route, and 2) duopoly or oligopoly, when specific shippers and consignees are linked by two or more parallel railways, each of which can provide single line service, and/or participate in the supply of through interline service. However, these simplistic renditions of market structure do not reflect other conditions which affect the nature of intramodal rail competition in particular situations, i.e.,: 1) the different ways in which shippers' selections of particular routes can be affected by a) reciprocal switching agreements at origin and destination points, b) transit privileges, c) carriers' car supply capabilities, and d) track and terminal conditions, circuity, and other aspects of the railway operating environment; 2) market and production point competition, which affects movements to and from particular points; 3) monopsony and oligopsony in certain large freight markets; and 4) shippers' and consignees' locational elasticities. In short, the structure of intramodal railway markets for carload and trainload traffic is a complex mixture of varying degrees of competitiveness shaped by numerous distinct conditions which vary among specific rail freight transport markets and time periods. Furthermore, the complexity of this competitive milieu is compounded by intramodal rail competition involving *TOFC/COFC* traffic, and intermodal competition involving both carload and *TOFC/COFC* service.

II. Limits on Intramodal Rail Rate Competition

Pervasive and intense rate competition is not feasible under the aforementioned structural conditions. Vigorous rate rivalry, with costs traceable to specific movements on a causal basis serving as floors for the pricing of individual movements in a significant number of markets or in those markets which generate great portions of revenue, will drive the general level of rates and hence total revenue below the sum of divisible and indivisible costs. Railway companies involved in such a disequilibrium and desirous of behaving rationally could, individually, achieve only short-run gains by accepting traffic at rates equal to or greater than traceable (marginal) costs but less than average costs—until their longer-lived assets become fully consumed, at which points in time curtailments of service and outright abandonments would be forced, absent 1) recourse to collective rate increases, or 2) external subsidy of shortfalls between total revenues and aggregate (divisible and indivisible) costs.

This pattern of events preceded the emergence of collective ratemaking in the nineteenth century. Such ratemaking did not stem wholly from a desire to generate monopoly profit. Rather, it was stimulated by imperfections in competitive dynamics; unilateral rate increases and/or disinvestments could not serve to bring about satisfactory rate-cost equilibria. The railway industry, viewed historically, exhibited conditions identified with unstable duopoly and oligopoly, in which the phenomenon of price leadership associated with relatively stable oligopolistic industries would not serve to prevent or end prolonged or recurrent outbreaks of ruinous rate competition.

It can be hypothesized that the tendencies toward such instability are minimal in the contemporary railway industry, and will decline further if recently announced merger proposals carry through and produce an industry populated by approximately six major railways and a larger number of smaller-sized regional and local carriers. Space precludes further pursuit of this hypothesis here; in any event, it cannot dispel the indeterminacy of specific outcomes in intramodal railway competitive relationships. The probability of an outbreak of intense, pervasive intramodal railway rate competition (with traceable marginal costs as pricing floors) might be relatively low, but the inherent incompatibility between such competition and the achievement of appropriate aggregate price-cost relationships in rail transport is a continuing reality.

The feasibility of intramodal rate competition is also limited by the railway industry's unique blend of rivalry and interdependence. Individual rail firms both connect with, and in many areas, parallel one another or serve the same origins and destinations. This phenomenon, together with the fact that many shippers and receivers of carload freight have private siding access to only one carrier at particular sites, places two or more carriers in the position of both competing with and complementing one another on freight movements between the same pairs of origin and destination points. Thus, for example, a carload movement can be handled entirely over the line of either of two parallel railways *if* the consignor's *and* consignee's premises are entered by *both* carriers. However, if the consignor's siding is served by *one* of the parallel carriers, while the consignee's property is reached *only* by the other railway, fulfillment of the consignor's and consignee's carload service demands will require interchange between the parallel carriers. A mix of interdependence and rivalry also emerges between carriers which link the same market areas with diverse points of production for competing raw materials and finished goods, and between carriers which link the same production points with different markets.

Interdependence is of greater importance than rivalry in the railway industry's aggregate traffic base; more than half of all shipments pass over the rails of two or more carriers. The sale of interline freight services on an all-inclusive basis thus requires intercorporate cooperation and coordination of a type and scale unknown and unnecessary in other sectors of business. And, carriers which participate in the movement of interline traffic on an all-inclusive basis (i.e., a single rate quotation for the movement, uniform minimum shipment quantities, and other uniform price-related terms of sale) must share responsibility for all key aspects of the pricing process.

The continued existence of such joint economic decision making was dealt a grave if not fatal blow by the ICC's August 13, 1980 decision. Although the decision, at face, does not prohibit continued antitrust immunity for collective interline ratemaking, it defines carriers who can practicably participate in interline traffic as "...only those carriers who are *direct connectors* to a specific joint-line movement of a specific commodity." This narrow circumscription of the range of continued antitrust immunity will make illegal and thus terminate the construction of tariffs with comprehensive coverage of commodity types, alternative routes and breadth of geographic service area, and available carriers.

The ICC's decision also requires that rate bureaus' rate committee meetings (under the sliver of immunity which will remain if the decision holds) be entirely open to the public, and that sound recordings be kept of each meeting. Fulfillment of this requirement will open all dimensions of rail carriers' interline pricing strategy to competing transport media, and to shippers, some of whom possess monopsony and oligopsony power and wield it relentlessly in quest of short-run marginal cost-oriented rates. It will also expose business information concerning freight movements which traditionally has been provided in confidence by individual shippers seeking new or adjusted rates. Such conditions obviously will force railways to abandon rate bureau-based interline ratemaking and shield their competitive advantages by using other means for the pricing of interchange traffic.

One alternative is the construction of joint through rates on an individual, carrier-by-carrier, movement-by-movement basis. However, this approach would, in comparison with existing rate bureau-based procedures, entail greater administrative effort and cost for carriers, and lengthen the time required for establishing and adjusting equivalent numbers of interline rates. A multiple of points of contact would have to be substituted for the one or relatively few which are presently provided to both carriers and shippers by rate bureaus, with concomitant rises in communications, record keeping, and travel expenses.

The use of such an atomized approach to interline pricing, together with the prohibition of both voting and discussion on single line rates, also increases the possibility of rate actions which could be in violation of the Interstate Commerce Act's proscriptions of unjust discrimination, undue preference and prejudice, and long haul-short haul discrimination. The ICC admitted the existence of this danger in its August 13 decision, but asserted that its intended encouragement of competitive pricing would make it "...unlikely that discrimination or preference will be claimed in the future merely because single-line and joint-line rates are different." The ICC went on to express the view that it would be more desirable to "...let aggrieved parties seek appropriate administrative or judicial remedies..." if unlawful discrimination or preference do occur, rather than to prevent prosecutable missteps via collective action.

Faced with such risk of prosecution and its substantial cost, railways might withdraw altogether from the promotion and sale of interline services, thus eliminating joint through rates, together with through bills of lading and single freight bills for multiple-line hauls. Shippers requiring interline rail service would be forced to deal separately with each carrier needed for provision of the service. Since the number of carriers participating in a joint movement can involve three, four, or even more, the transaction expense element of rail service purchases would rise. Shippers would have to expand their tariff files, computer-based information systems, and personnel required for the procurement of pricing and service information and the settlement of freight bills on interline consignments, because only single line tariffs would be available, and separate billings would have to be made by each of the carriers participating in an interline movement. Shippers would also bear responsibility for resolving all terms of service (for example, commodity descriptions, weight minima, equipment specifications, and packaging requirements) individually with each carrier. Some individuals have cited recently initiated interline rail movements of regulation-exempt agricultural products as demonstrating the effectiveness of joint ratemaking *sans* antitrust immunity and rate bureaus. However, such movements involve an extremely limited number of routes, carriers, origins, and destinations, and thus cannot serve as a model for the pricing of consignments moving between a multitude of origin and destination points over numerous interline routes.

The increased transaction cost, time, and complexity occasioned by the demise of tariffs for geographically comprehensive interline service would decrease railways' attractiveness vis-à-vis single line service by carriers in other modes. Traffic loss potential would be greatest for the carload segment of the national rail freight traffic base (as distinguished from the *TOFC/COFC* and unit train categories), and for short-line railways lacking in pricing and marketing capabilities and bargaining power with larger connecting carriers. One result of this might be intensification of the railway industry's drive toward corporate consolidation, so as to broaden the areas over which individual systems could offer single line service—a startling irony, given the ICC's new emphasis upon intramodal competition.

III. Cartelism vs. Contemporary Collective Ratemaking

The railway industry, judged in terms of return on net assets, does not exhibit monopoly profits. Nevertheless, contemporary rate bureau-based pricing is portrayed as cartelism by its opponents. They allege that the bureaus serve principally to maintain misallocative value-of-service rates based upon floors that reflect the higher operating costs of the bureaus' least efficient members. They thus indict rate bureaus for 1) occasioning welfare losses via excessively priced and priced-out traffics; 2) inhibiting innovation; 3) causing inefficient location of economic activities; and 4) maintaining identical rates over large numbers of alternate through routes within the national railway network, thus inhibiting a) concentration of traffic upon lowest-cost, least-circuitous routes, and b) abandonment of redundant high-cost mainline route mileage.

Conditions relating to contemporary rate bureau operations do not sustain these accusations. To begin with, the right of independent action by individual bureau members is specified and protected by law, and independent actions on particular rates are frequent and numerous. Information obtained by the writer through more than a decade of discussion with a variety of individuals involved in the processes of different rate bureaus indicates that the threat of independent action also influences the outcomes of collective determinations of both particular rates and general rate levels. And, contrary to the speculative induction of certain economists, such collective determinations are not governed by the cost levels of inefficient carriers. Rather, they are the result of a complex of determinants, including rate and cost levels of competing modes of transport, competition between different market and production areas and among substitutable commodities, and the bargaining skills of carrier and shipper executives. Consideration of shippers' views is an integral part of rate bureaus' processing of proposals for new and adjusted rates on particular traffics, and law requires that shippers be given adequate notice and opportunity to comment on collectively determined general rate change proposals before they are submitted to the ICC. Such buyers' participation does not comport with cartel theory, nor does the fact that reductions dominate rate bureau actions on particular rates. In 1977, for example, the Western bureaus processed 8,724 rate change proposals, of which 7,633 (87 percent) were reductions. This poses the question of whether collectively determined reductions would be higher or lower if effected independently (assuming it to be feasible). Efforts to answer it would, however, become mired in indeterminancy; the presence of substantial cost indivisibilities limits the meaningfulness of estimates of losses and gains in shippers' welfare which are based on differences between 1) alternative levels of rates on particular units of traffic, and 2) the marginal (traceable) costs of those traffics. Shippers' welfare obviously will suffer if revenues on individual traffics, taken *en toto*, fall short of indivisible costs and force service withdrawals in instances where the aggregate uncaptured value of such service (to shippers) equals or exceeds the summation of carriers' divisible and indivisible costs.

The sluggish rationalization of purported excess mainline route capacity and uneconomic circuity in carload shipment routings cannot be linked largely or solely to collective pricing. Rather, these topics are bound up with other conditions, such as 1) regulation of abandonment and provision of through routes and joint rates; 2) the longevity and specialized nature of railway fixed plant, which prolongs the period within which operations can continue at revenue levels below those required to sustain major capital renewals; 3) ambiguity and misapplication of criteria for ascertaining excess fixed plant (i.e., underutilization as specified by differences between actual traffic volume and full traffic capacity of a particular route a) involves unsettled questions in physical capacity measurement, and b) is not a conclusive indicator of the capacity's value as measured by its contribution to a carrier's profitability), and 4) the need for a proper distinction between *line* and *route*; many lines with unused capacity exist primarily for the servicing of local traffic. Through traffic thus can often be accepted at relatively low marginal cost and will yield additional contributions toward coverage of fixed costs.

IV. Conclusions

Contemporary rate bureau-based railway ratemaking is not cartelism. Rather, it is a process of collective bargaining between carriers and shippers, involving interactions of intra- and intermodal competitive forces and rivalry between alternative production and market sites and substitutable commodities. This bargaining process provides a means for facilitating interrailway analysis and decision making, and exchanges of information between shippers and carriers, in the pricing and sale of interline traffic on an all-inclusive basis. If the ability to do so should be lost through removal of antitrust immunity, transaction costs for sellers and buyers of interline railway freight services will rise, and such services will decline in competitive advantage vis-à-vis the services of carriers in other modes capable of providing single line movements. The railway industry will be balkanized further in terms of its ability to act as a nationwide system, particularly in the handling of carload (as distinguished from *TOFC/COFC* and unit train) consignments. It also appears likely that the industry will be exposed to risks of litigation stemming from ambiguities in the applicability of various Interstate Commerce Act and/or antitrust provisions.

The general potential for welfare losses from elimination of useable rate bureau antitrust immunity hence is clearly identifiable. In contrast, prospects for welfare gains from subjection of railway pricing to antitrust standards are nil, given the railway industry's inherent cost conditions and other structural traits which bar the feasibility of pervasive and intense intramodal rate competition and require individual railway firms to function as an integrated system for the purpose of providing geographically comprehensive service in competition with other forms of transport. Thus, net cost, rather than gain, will ensue if the current tide of oppostion to collective railway ratemaking is not reversed.

The Nature, Effectiveness, and Importance of Motor Common Carrier Service Obligations

By BENJAMIN J. ALLEN*

The 1980's began in the midst of a debate over trucking deregulation and by the seventh month of the decade, the Motor Carrier Act of 1980 became law. At the heart of this debate over trucking deregulation is the issue of the importance of motor common carrier service to the viability of small rural communities. Under regulation, motor common carriers benefit by being protected against potential new entrants in their markets; but in return, they have a common carrier obligation to provide service to all communities for which they have authority, although some communities may be unprofitable or less profitable than other markets. Opponents of deregulation of the interstate trucking industry argue that deregulation would terminate common carrier service obligations and thereby lead to the cessation or at least a deterioration of for-hire motor carrier service at "reasonable rates" to small isolated communities. On the other hand, proponents of trucking deregulation challenge the notion that the motor common carrier system actually requires cross subsidization of service to small communities with profits generated on higher volume routes. The proponents argue that such practice would be irrational for the profit-oriented business firm and that it is unrealistic to think that the Interstate Commerce Commission (ICC) can oversee and enforce the service obligations of some 12,000 motor common carriers. This paper addresses this continuing debate by analyzing the role and importance of the motor common carrier service obligations in ensuring adequate trucking service to small rural communities. In addition, the possible role of the motor common carrier service obligations in "protecting" small communities in the 1980's is discussed.

I. The Nature of Motor Common Carrier Service Obligations

The degree of "protection" afforded to small communities by the motor common service obligations depends upon both the nature of the obligations and the effectiveness of their enforcement. The important public policy question is, "Do the service obligations, if enforced, require more of motor common carriers than what unregulated profit maximizers would tend to offer?" If not, the adequacy of service in small communities now attributed to economic regulation actually results from profit-maximizing behavior of the firms, not from the enforcement of the common carrier service obligations.

In a recent study, Denis Breen and I clarified the nature of the service obligations of regular-route motor common carriers by delineating the dimensions of service which are dictated by the common carrier obligations and those which are not. We concluded that motor common carriers must hold themselves out to serve the general public indiscriminantly up to the limits of their operating authorities. Thus, motor common carriers are required to "hold themselves out" to serve all communities in their operating authority, including small rural communities. That common carriers are permitted little discretion as to *who* is served, that is, traffic selectivity is prohibited, implies that more is required of these carriers than what unregulated profit max-

*Associate professor, School of Business Administration, Iowa State University. I would like to thank William H. Thompson, David B. Vellenga, and Roy D. Voorhees for their comments on earlier versions of this paper. I am particularly indebted to Kevin Boberg for research assistance and comments on an earlier version, and Denis A. Breen, who provided helpful suggestions on this paper and initiated many of the research ideas associated with this topic.

imizers would tend to offer. At the same time, however, we found that much managerial discretion is permitted with respect to the *quantity* and *quality* of service a common carrier may offer. The common carrier service obligations have been interpreted to permit wide managerial discretion with respect to frequency of service, provision of specialized equipment, and maintenance of excess capacity to meet unexpected demand. Common carrier behavior implied by the service obligation is indistinguishable from unregulated profit-maximizing behavior with respect to these dimensions of service. Overall then, it appears that less is required of motor common carriers than what the conventional wisdom, as seen in recent shipper testimony before Congress, would indicate.

Although the ICC attempted in the 1970's to reduce the amount of the traffic selectivity practiced by motor common carriers, it also took steps which weakened the common carrier service obligations. For example, the ICC weakened the obligations with respect to the practice of a motor common carrier interlining instead of providing direct service at small communities which the carrier is authorized to serve. Prior to 1970, the ICC had usually ruled that this practice violated the carrier's common carrier service obligations and had required the carrier to reinstitute direct, single line service to small communities. In its decision in the Consolidated Freightways pooling case in 1971, the ICC changed its policy by taking a favorable stance toward pooling agreements. The ICC ruled in the Consolidated Freightways case and subsequent cases that interlining to authorized points is consistent with the common carrier duty to serve if done under an approved pooling agreement. The common carrier obligations of a carrier were also weakened in the late 1970's by the ICC's recognition and sanction of the use of convenience interlining. The "convenience interlining" rule allows a carrier, through a simple tariff publication, to provide through service to its authorized points by interlining at the gateways or interchanges points where joint through routes are applicable. Insufficient time has elapsed to determine the effect of this new policy on service to small rural communities although the pooling cases indicate that trucking services can actually improve under the pooling agreement.

II. Effectiveness of the Service Obligations

Given the nature of the motor common service obligations, the degree to which the elimination of them under deregulation will lead to a deterioration of for-hire trucking service in small rural communities depends upon how effective the service obligations are under the present regulations in keeping common carriers serving unprofitable and less profitable rural communities. As noted, the common carrier service obligation may require a trucking firm, by forcing it to provide service to all communities listed in its certificate, to behave at times in a manner inconsistent with profit maximization. To expect a firm to serve a rural community which it finds to be unprofitable, one must assume that either the trucking firm will behave irrationally or that ICC regulation provides financial incentives for carriers to serve these markets through a cross-subsidy mechanism, or that the ICC forces the firm to serve these markets through regulatory coercion, that is, enforcement of the service obligations. Assuming away irrational behavior on the part of trucking executives, one is left with cross subsidy and regulatory coercion as explanations for carriers serving unprofitable communities. Michael Pustay, John Drake, and James Frew, however, recently examined the cross-subsidization mechanism for inducing motor carriers to furnish unprofitable small community service and concluded that certain variations of the mechanism cannot be applied to the trucking industry and that other "workable" variations have not been employed by the ICC.

Pustay, Drake, and Frew conclude that coercion is the only viable mechanism available to the ICC to ensure that carriers will offer service to small communities that are unprofitable or less profitable than their other markets. For common carrier obligations to be effective in these cases, they must be enforced by the ICC. There are statutory provisions establishing enforcement procedures, and the ICC does attempt to detect, investigate, and punish offenders. But the existence of statutory provisions and enforcement activities of the ICC does not guarantee that carriers will comply with their duty to serve. In fact, it is reasonable to believe that, given its other regulatory responsibilities, the ICC could not adequately monitor the behavior of 12,000 interstate motor common carriers. Darius Gaskins, Chairman of the ICC, acknowledged that the ICC does not enforce the common carrier obligation and, furthermore, does not have the information needed to enforce it. Thus, the common carrier obligation to serve small communities may be illusory unless the shipper initiates a complaint against a carrier and follows through with it at the ICC. Shippers in rural communities, however, are generally unaware of ICC enforcement procedures and rarely rely upon them when service problems do arise (see, for example, William Thompson, Roy Voorhees, and Kevin Boberg).

The degree to which the service obligation forces a carrier to serve unprofitable small communities not only depends upon the probability of a service failure being detected by the ICC or reported but also on the severity of the penalty. Penalties range in severity from letters and phone calls from the ICC to the offending carriers to suspension or revocation of the carrier's certificate —a penalty which has rarely been employed. The ICC can also seek injunctive relief in district court and can use a record of past violations of the common carrier service obligations as a basis for finding a carrier seeking additional operating authority to be unfit.

Several studies have been recently conducted to determine how effective the regulation of the service obligation has been in ensuring that motor common carriers are actually offering service to all of their authorized points—including small rural communities. In our study in the Pacific Northwest, Breen and I found a substantial amount of service failure by fifteen motor common carriers. These carriers tended to abandon communities in declining counties and generally failed to offer service at all authorized points acquired through certificate purchases. Furthermore, most communities were not being served by all their authorized carriers and some communities were not being served by any of the major authorized long-haul carriers. Similar findings were obtained in other studies including a group of community case studies conducted by the U.S. Department of Transportation (for example, see Karen Borlaug et al.), a study by Paul McElhiney in Wyoming, and a study by Policy and Management Associates, Inc. which surveyed more than 300 motor common carriers.

III. The Necessity of the Service Obligations

The regulatory mechanism both defended and doubted as ensuring that a carrier will offer service at its less attractive markets is the regulation of the motor common carrier service obligation. Thus far, it has been shown that the nature of the common carrier service obligation does not control many dimensions of service which are important to shippers in small communities and that the ICC enforcement program has been largely ineffective in enforcing the one dimension of service obligations address— traffic selectivity.

If the analysis stopped after examining only the nature and effectiveness of the motor common carrier service obligations, inappropriate policy recommendations might be made. For example, given the above information some might recommend strengthening the service obligations, for example, by eliminating managerial discretion with respect to frequency of service or by devoting more resources to enforcement activities. The analysis must be extended to an examination of the question, "Are motor common carrier service obligations necessary for ensuring adequate trucking service to small communities?"

There are two issues pertinent to this question: 1) Given the apparent widespread service failure by motor common carriers, how adequate is trucking service offered to small communities; and 2) is small community trucking service profitable?

A. *Adequacy of Existing Trucking Service*

If evidence is found that small communities are not receiving adequate transportation because of the failure to keep many carriers in small communities, a case for strengthening the enforcement of service obligations might be made. Several recent studies have found small communities were receiving adequate trucking service despite the fact that many of their authorized carriers were not offering any local service (See Breen and Allen, and Borlaug et al.). Most shippers/receivers in small communities that were surveyed in these studies thought their overall level of general commodity transport service was adequate when all sources of supply (large interstate carriers, small short-haul interstate carriers, United Parcel Service (UPS), and private carriage) were considered. The motor common carrier service to the small communities was predominantly offered by the smaller (class II and class III carriers) short-haul carriers. Although its services are limited, UPS was considered by most shippers and receivers to be offering an essential and superior service to that provided by the regulated route common carrier. The overall adequacy of service depended more on the actions taken by the shippers and less on the regulation of the service obligation, given the evidence of widespread service failures in these communities. The tendency for shippers to make market adjustments, for example, switch carriers and/or types of carriers, to service problems instead of using the ICC regulatory apparatus was also found by Thompson, Voorhees, and Boberg.

The available evidence indicates that the small rural communities are receiving adequate trucking service despite service failures by many authorized carriers. Interestingly, many of the complaints about the adequacy of common carrier trucking service in small communities were about dimensions of service, for example, frequency of service, that are not controlled by the service obligations or any other ICC regulation. The results of these studies do suggest that one should treat with skepticism the usual association made between the regulation of the service obligation and the finding of adequate trucking service in small communities.

B. *Profitability of Small Community Service*

Although the recent studies indicate shippers/receivers in small communities depend upon various sources of trucking supply, many do rely upon regular-route motor common carriers of general freight because they ship or receive packages too large for UPS, or because private trucking is not feasible. Thus, it is necessary to examine whether the regular-route common carriers actually providing service will continue to do so in a deregulated environment. If the traffic is profitable for those firms actually serving, the service is likely to continue if and when the common carrier obligations are terminated.

Although small communities have traffic characteristics which tend to make them more costly to serve, the unattractiveness of the rural traffic might be overstated because carriers specialize in providing service to small communities. For example, in a study by R. L. Banks and Associates of nine motor common carriers specializing in serving small, rural communities, it was found that small class I and class II carriers succeed because they appear to be better equipped to handle traffic in small markets due to the fact that their pick-up and delivery service, as well as their terminal operations, are geared for small LTL shipments. In addition, providing service to rural communities can also be made profitable by adjusting rates on shipments to and from such communities. A major finding of a study by Paul McElhiney of the freight rate structure throughout a nine-state western region was that small shippers in rural communities in this part of the country are often subject to through rates plus arbitraries of local rates. Based on the 1976 Continuous Traffic Survey, as much as 13 percent of all truck shipments in towns of 1,000 to 5,000 persons involve arbitraries. (See Congressional Budget Office.)

Pustay, Drake, and Frew examined the profitability of small community trucking services. Although they could not conclude that ICC regulation had no effect on extending service to small communities, they did conclude that in the main, apparently the existing small community services are furnished because such services are profitable for the carriers to supply.

IV. Common Carrier Service Obligations—Their Role in the 1980's

When Congress was drafting the Motor Carrier Act of 1980, it could have taken one of two basic but opposite approaches to ensure adequate motor common carrier trucking service to small rural communities in the 1980's. One approach is to maintain the existing regulatory scheme but require the ICC to enforce the service obligations of motor common carriers. The other approach would be to eliminate the common carrier service obligations through deregulation and depend upon competitive pricing and free entry to "protect" shippers in small communities. The various versions of the bills leading up to the Motor Carrier Act of 1980 contained elements of both of these approaches. It is not clear what the congressional intent is in the Motor Carrier Act with respect to the relative roles of increased competition and the regulation of the common carrier service obligations in ensuring adequate regulated trucking service to small communities. On one hand, Congress added to the National Transportation Policy a phrase indicating a new policy goal of promoting competitive and efficient transportation services in order to provide and maintain service to small communities and small shippers. On the other hand, Congress instructed the ICC to study the common carrier obligation to provide service to small communities and to assess the enforcement of the obligation. If the ICC finds it is not fully enforcing the obligation, it is required to explain why it has not and to estimate the resources that would be required for full enforcement. In short, it would appear that Congress sidestepped the issue, at least temporarily, and decided to have the ICC continue to "set policy" with respect to small community service. In recent years, the ICC's policy has been to gradually reduce the role of the common carrier service obligation and increase the role of competition in ensuring adequate trucking service in small communities. As mentioned above, by permitting the use of pooling agreements and convenience interlining, the ICC has made at least one dimension of the common carrier obligation less constraining through administrative action. The increasing role of competition can be seen from the fact that the ICC has also relaxed entry into small, rural markets and evidence exists that numerous applications for such authority have been granted. (See Congressional Budget Office.)

V. Conclusion

The late 1970's produced a number of studies indicating generally that ICC regulation of the common carrier service obligations is neither capable of nor necessary for ensuring adequate trucking service to small rural communities. Even if some "protection" of small communities were necessary, a more desirable approach in terms of equity and efficiency would be to use direct subsidy—not cross subsidy. Congress had an opportunity at the beginning of the 1980's to establish a new policy with respect to ensuring adequate trucking service to small communities but chose instead to request additional studies and to maintain the status quo. One can thus safely predict that the small community and motor common carrier cross-subsidy argument will be dusted off and used again by opponents of deregulation when the ICC makes its legislative recommendations based on its study of the common carrier service obligation to provide service to small communities.

REFERENCES

R. L. Banks & Associates, Inc., *Economic Analysis and Regulatory Implications of Motor Common Carrier Service to Predominantly Small Communities*, final report to the

U.S. Department of Transportation, June 1976.

Karen Borlaug et al., *A Study of Trucking Services in Six Rural Communities,* U.S. Department of Transportation, Nov. 1979.

Denis Breen and Benjamin Allen, *Common Carrier Obligations and the Provision of Motor Carrier Service to Small Rural Communities,* final report to the U.S. Department of Transportation, July 1979.

D. Gaskins, "Comments," Meet the Regulators Program, American Enterprise Institute, Washington, Mar. 11, 1980.

Paul McElhiney, *Motor Common Carrier Freight Rate Study,* prepared for Federation of Rocky Mountain States, Inc., Denver, May 1975.

Michael Pustay, John Drake, and James Frew, *The Impact of Federal Trucking Regulation on Service to Small Communities,* final report to the U.S. Department of Transportation, Mar. 1979.

William Thompson, Roy Voorhees, and Kevin Boberg, *Motor Carrier Service to Small Communities in Iowa,* report prepared for Iowa Department of Transportation, May 1980.

Congressional Budget Office, "The Impact of Trucking Deregulation on Small Communities: A Review of Recent Studies," staff working paper, Natural Resources and Commerce Division, Feb. 1980.

Interstate Commerce Commission, *Consolidated Freightways Corporation of Delaware, et al, Pooling,* 109 M.C.C.596 (1971).

Policy and Management Associates, Inc., *The Impact on Small Communities of Motor Carrier Regulatory Revision,* prepared for U.S. Senate Committee on Commerce, Science and Transportation, June 1978.

The Role of the Interstate Commerce Commission in the 1980's

By JOHN GUANDOLO*

If recent experience is any measure, it appears that in the 1980's the Interstate Commerce Commission will be its own master, charging forward to change the way the Interstate Commerce Act is administered, then awaiting ratification by Congress. In making these types of changes recently, the ICC's guiding light has been the effect an action will have on competition in the marketplace. If competition among carriers is increased, then the ICC finds action under consideration is necessarily good, and implements it. At the same time there is a contradictory impulse, particularly in the regulation of rail carriers, to carve out an area within which revenue is protected, such as by immunizing their rates from attack by shippers on grounds that the rates are too high, or by encouraging contract rates. It is the interplay between these two conflicting impulses, one towards competition and the other towards stabilizing and preserving revenue, which will determine the course that the ICC follows in the 1980's.

In addition, the ICC has shown a tendency towards abdicating many of its regulatory functions and relying on the industry for self-policing. While this is a significant departure from many government agencies' drive to amass power, it is nevertheless real and must be reckoned with in the coming decade.

Perhaps it was in the *Liberty Trucking Company, Extension—General Commodities* case that the ICC most clearly embraced the proposition that "competition is generally presumed to be in the public interest." This case was part of a series of cases relaxing the requirements for entry into the trucking industry, clearly favoring the carrier who wished to enter a market over those that were already serving that same market. In a later *Policy Statement on Motor Carrier Regulation*, the ICC removed the question of whether the carriers presently serving the market could themselves provide the service as a factor in determining whether the authority sought should be granted. This in turn was the forerunner of the recently enacted Motor Carrier Act of 1980. The Motor Carrier Act ratified the ICC's earlier liberalization of entry requirements, and created a new statutory standard. The requirement that a carrier desiring new operating authority must demonstrate that its operation will meet the "public convenience and necessity" was changed to a requirement "that the service proposed will serve a useful public purpose, responsive to a public demand or need" (49 U.S.C. §10922(b)(1)). While it is unclear how the requirement that a "public need or demand" be shown favors competition more than did the old standard, "public convenience and necessity," it is not open to doubt that it will be used by the ICC to grant new authority wherever this action can be construed as increasing competition.

What, however, will be the effect of easier entry? In an ideal world, many new motor carriers would file applications and begin operations, challenging the giants of the industry to lower rates and improve services. Indeed, to some extent this should be encouraged by a number of specific provisions in the Motor Carrier Act, such as those that allow owner-operators to transport food and fertilizers along with products exempt from regulation upon a showing of fitness, and that provision which exempts from ICC regulation line-haul transportation which is incidental to air transportation. However, while the jury is not yet in on the overall effects of relaxed entry on competition, there

*Partner in the firm of Macdonald & McInerny, P.C. In preparing this article, I was assisted by John Downing, an associate with the same firm.

is the disturbing possibility that the result may be to encourage the formation of limited numbers of nationwide carriers who will handle the bulk of the available traffic. There has already been some sign that the new legislation encourages this trend as carriers are filing applications for territory-wide nationwide, general commodity authority, specifically relying on the new act.

Another effect of relaxed entry criteria which will play an increased role in the 1980's is the likely diminution in value of motor carriers' operating rights, which are now carried on their books as assets, and are often pledged as security in order to obtain financing. With the increased ease of obtaining operating rights by filing applications with the ICC will come a decrease in the number of carriers wishing to purchase those rights from other carriers. The probable decrease in value provided a major impetus for opposing the legislation for many carriers. Probably so as to allay the motor carrier industry's fears that certificates will become worthless, Congress forbade the ICC to use the master certification approach it had proposed in Ex Parte No. MC-135, *Master Certificates and Permits*. Under this approach the ICC would have made a general finding of public convenience and necessity for service of certain specialized types, such as service by heavy haulers or by household goods carriers. Thus, upon application by a carrier desiring to provide the service, and after satisfying the ICC that it was fit, the carrier would be notified to begin operations. While this approach has been eliminated by Section 5 of the Motor Carrier Act, 49 U.S.C. §10922(b)(2), it is doubtful that the ICC will be greatly hindered in removing entry barriers, since it had been granting more than 95 percent of the operating rights applications filed even prior to the legislative revision of the "public convenience and necessity" standard.

Other means by which competition is meant to be encouraged include the prohibition of use of the rate bureau procedure for discussion of single line rates starting in 1984 (49 U.S.C. §10706(b)). Since 1948 the ICC has had power to immunize motor carriers from the operation of the antitrust laws in order to carry out their function of setting rates. The new legislation represents the recognition by the Congress that there is no compelling public interest in allowing groups of carriers to set rates which affect movements involving only one carrier, and thus do not require joint discussion and consideration. However, the legislation allows groups of motor carriers to continue to consider general rate increases and tariff restructuring, though recognizing that each of these will involve changes in single line rates. However, the information considered in debating such matters is strictly limited to industry average costs, and may not include discussion of specific markets or single line rates.

The intent of this legislation is clearly stated in the House Committee's report:

> It will put a greater burden upon individual carriers to determine their own cost structures and the most optimum rates from their individual company point of view to offer the shipping public. But on balance, the Committee believes that there will be public benefits that outweigh the burdens of changing to a new system—namely, more competitive pricing. [p. 5]

It is expected that the Commission will implement this provision as swiftly as the law allows, since it had earlier announced its decision to eliminate rate bureau immunity for single line rate setting under its then-existing powers.

In contrast to the direction that competition be increased, with the corollary that rates will tend to a lower level, the legislation has made it easier for motor carriers to raise rates. In Section 11 of the Motor Carrier Act of 1980, 49 U.S.C. §10708, individual motor carriers and freight forwarders are given complete freedom to raise (or lower) rates so long as they are not changed by more than 10 percent above (or below) the level one year prior to a specified point of reference. The ICC can increase the zone to 15 percent for rate increases in any year that it finds that there is sufficient competition to regulate rates, and that the carriers, shippers, and the public will derive a benefit from the increased flexibility. This provision preserves rates within that range from even the chance of attack by shippers on grounds that they are too high, or by competing carriers claiming that they are too low.

While the House Committee report acknowledges that this rate freedom provision is intended as a balance against freer entry, freer expansion of the industry, and increased competition that results from other portions of the legislation, it should be remembered that it applies only to individual carrier rates, and thus does not generally apply to across-the-board rate increases. At this writing, the ICC still has under consideration whether to adopt standards which would place an upper limit on the profitability of any group of motor carriers proposing a general increase (Ex Parte No. MC-128, *Revenue Need Standards in Motor Carrier General Increase Proceedings*). Among the matters it is considering is whether to dictate an industry-wide standard rate of return based on investment or on equity. The legislation speaks only in general fashion to this issue stating that the ICC shall authorize revenue levels sufficient to return operating expenses plus "a reasonable profit." Section 12 of the Act also makes general increases easier to obtain in one respect because it allows motor carriers to justify rate increases by "reasonable estimated or foreseeable future costs."

In short, the regulation of motor carriers is likely to be heavily influenced in the future by the ICC's desire to increase competition by making entry into the market easier. At the same time, the conflicting impulse to preserve motor carrier revenues and to reject master certification and maintain some semblance of the application process will act as a restraint on the tendency to let the market regulate itself.

Insofar as the railroads are concerned, future regulation was mapped out in the 4R Act. The declared policies of the 4R Act include balancing the needs of railroads, shippers, and the public; fostering competition among all carriers in order to promote more adequate and efficient transportation services and the attractiveness of rail investment; permitting greater railroad price flexibility for rail services in competitive markets; promoting a rate structure more sensitive to variations in demand and separate rates for distinct services; formulating standards and guidelines for determining adequate rail levels; and modernizing and clarifying the functions of railroad rate bureaus. The changes in the 4R Act are intended to inaugurate a new era of competitive pricing among rail carriers. Proposed legislation also is stressing greater railroad price flexibility.

During the coming years, the Commission's regulation of railroads will probably experience its greatest changes in the rate area. The ICC views the railroads as a depressed industry, since many carriers have a relatively low rate of return. Short of offering the railroads millions of dollars of additional federal funds, the only means of increasing their income is to allow them to charge higher rates to those shippers who have no effective alternative to railroad transportation. Obviously, if there is an effective alternative, the shipper will simply switch his traffic from the railroad to the alternative mode, creating a loss of funds to the railroads. This puts a great burden on the shipper who has no alternative means of transportation, and sets up the problem of whether a rate increase can be lawful where the former rate already returned a reasonable amount, but the carriers' overall revenues are inadequate.

The ICC's apparent answer is that such a rate is lawful. The ICC has devised a variety of ways in which amounts above the full cost of the captive movements can be recovered. Up until recently, the ICC has allowed a 7 percent additive above fully allocated costs on movements of coal, the railroads' primary captive commodity. In June, a federal court remanded a case to the ICC for explanation of the choice of a 7 percent additive, stating it could discern no rationale for that figure (*Burlington Northern, Inc.* v. *United States*).

While this may prove a temporary setback in the ICC's effort to assure the railroads adequate revenue at the shippers' expense, it is unlikely that the ICC will fail to devise some rationale or substitute, so as to continue its policy. In the meantime, the likely direction of the ICC's future action is indicated by other proposals now under consideration. Under current law, the ICC can only inquire into whether a railroad's rate is unreasonably high after it is shown that the rail carrier has market dominance over the traffic. Market dominance, which is defined by statute as the absence of effective competition, is demonstrated by meeting any of the following tests: 1) where the carrier has handled 70 percent of the traffic during the preceding year; 2) where the rate exceeds the variable cost of the service by 60 percent or more; and 3) where the shippers have made a substantial investment in rail equipment or facilities.

In *Rail Market Dominance and Related Considerations*, the ICC has proposed to do away with its market share and substantial investment tests, leaving the variable cost test as the only sure way of establishing market dominance. However, the ICC has proposed to change this test as well. Where previously the reasonableness of the rate became an issue whenever a revenue/variable cost ratio of 160 percent was shown, the ICC proposes to raise that level to 180 percent. Rates returning revenue between 150 and 180 percent of variable cost would not raise any presumption of market dominance so that rate could not be inquired into unless market dominance was shown by other facts on a case-by-case basis. Rates returning less than 150 percent of variable cost would be presumed not to be market dominant, and their reasonableness could not be inquired into. Thus, the ICC's aim is to create a class of rates which cannot be challenged, to restrict the number of rates which can be challenged under its presumption, and to leave to its own discretion all those that fall in the middle.

In two rule-making proposals dealing with general rate increases, there have also been unmistakable signs that the ICC wishes to allow the railroads the certainty that many of their rates will not, and in fact cannot, be challenged on grounds that they are too high. In *Railroad Cost Recovery Procedures*, the ICC proposed a zone of reasonableness based on the average increase in all railroad costs within which rate increases could not be challenged. What the ICC apparently intends to accomplish is the abolition of the nationwide, largely uniform general increase which the railroads have traditionally used. By allowing individual carriers to elect an increase up to the amount needed to recover all cost increases, without challenge, the ICC hopes to allow individual carriers to recover their common cost increases, without having them do so as a group.

In *Modification of Rail Carrier General Increase Procedures*, the ICC proposed to adjust its requirements for submission of data in general increase proceedings, deleting among other items, the requirement that specific cost and revenue data, by commodity, be submitted. In effect, this deletion will remove the general rate increases on particular commodities from challenge. There will no longer be data available by which a shipper opposing the increase on a specific commodity can demonstrate that too high an increase is being sought. This is of particular importance when it is realized that rates on certain commodities, such as coal, are likely to increase far more than will rates as a whole because coal shippers generally have no alternative but to use the nearest available railroad for the high-volume shipments they must move.

While the three proceedings just mentioned foretell the ICC's future direction in its regulation of railroad rates, that is, increased immunity from attack with the probable consequence of heavy rate increases on captive traffic, the proposed Rail Act of 1980, considered in the House of Representatives, presents the clearest picture of the trend in this direction. The proposed legislation affords a rare glimpse of what measures the ICC may take in the future, since before its amendment the Commission had given it strong support.

Prior to its amendment, Section 201 of the Rail Act of 1980 provided that a rail carrier may establish any rate it desires unless the ICC determines there is no effective competition with respect to the particular transportation service involved. While this in itself does not change the existing approach, Section 202 of the bill provides that there is effective competition for any movement for which the rate to variable cost ratio is no greater than the amount needed to receive all fixed or variable costs for a sample of carload movements. The figure obtained would then be applied without regard to specific movement or commodity to which the particular rate applied. The effect will be to authorize an open season on rates on commodities which, because of economies associated with high volume movements, have relatively low variable costs, since they will be measured against the higher industry average.

Another major development aimed at increasing the latitude allowed in individual rate adjustment is the emphasis the ICC has given to contract rates. In *Change of Policy, Railroad Contract Rates*, the ICC encouraged the filing of contract rates by rail carriers in appropriate circumstances. The decision made it clear that rail contract rates were not considered to be illegal per se, contrary to dicta in certain earlier proceedings. The ICC believes that negotiated ratemaking offers great promise for improved service, reasonably stable and predictable rate levels, and needed innovative rail pricing. Apparently ignored by the ICC are questions concerning the harm and injury that may be inflicted on competing carriers by traffic being tied up under the contract for a fixed period of time. In *Change of Policy, Railroad Contract Rates*, the ICC is now giving consideration to adopting standards and procedures for contract rates. In the alternative, the ICC is considering whether some or all aspects of the contract filing and regulation could be exempted from the ICC's jurisdiction. If regulation is retained, the rights and obligations of shippers and carriers concerning contract rates will be determined, and the ICC will adopt rules to remove existing barriers to the filing of these rates and encourage their use. In addition, the ICC has supported legislation removing contracts entirely from their jurisdiction, except for determination of whether a contract impairs the common carrier obligation.

A primary direction in rail regulation for the 1980's will be the continued diminution of the role of the railroad rate bureau. In 5b Application No. 2, *Western Railroads-Agreement*, the ICC has recently disapproved the rate-making agreements of the four major railroad rate bureaus because they allowed discussion of, and voting on, single line rates and rates by carriers which could not practicably participate in the movement. This action is plainly designed to encourage competition through individual rate setting. Just as plainly, it is a step on the way to making the rate bureaus primarily tariff publishing agencies. In this decade, we may well see the removal of carrier antitrust immunity for the setting of joint-line rates as well, a prospect which would increase the difficulty of setting new rates and responding to changing market conditions.

There has also been a strong recent trend toward the abdication of statutory authority by the ICC. The ICC has given the railroads flexibility in setting per diem below the applicable basic per diem set by the ICC and eliminated incentive per diem on boxcars and gondolas. See *Order Granting Railroads Flexibility in Setting Per Diem Rate Levels*; and *Elimination of Incentive Per Diem Charges*. In *Investigation of Adequacy of Railroad Freight Car Ownership, Car Utilization, Distribution Rules, and Practices*, the ICC repealed mandatory car service rules and has turned over the policing of car service to the industry. In *Proposal to Repeal Credit Regulations*, the ICC proposed abolition of its regulations on the extension of credit to shippers by carriers of all types, replacing them with a provision requiring only that credit be extended in a nondiscriminatory manner. The proposed Rail Act of 1980 also emphasizes the administrative reduction of regulation. It would replace the necessity of making a finding under 49 U.S.C. §10505 that rail regulation from which exemption is sought is unreasonably burdensome and serves no useful public purpose, with the simpler requirement that either the transportation or service involved is of limited scope, or that regulation is not needed to prevent abuse of market power. I believe we will see the exemption power used frequently to release the railroads from existing requirements such as car service orders, accounting regulations, and some tariff filing requirements.

In summary, I believe the 1980's will bring about increased competition in some areas by easing entry requirements and encouraging setting of individual rates. Simultaneously we will see a lessening of rate regulation, generally, which will favor the railroads and quite probably impose a greater burden on the captive shipper. The ICC appears ready, with the blessing of Congress, to plunge headlong into new areas both of regulation and of deregulation. It is opening itself to criticism seldom made of a government agency, that is, that it is moving too fast. The nation's transportation system has served the nation well. It is incumbent upon the ICC, therefore, not to make changes until it has a clear direction and is aware of the cumulative effect of the changes it advocates.

REFERENCES

Interstate Commerce Commission, *Change of Policy, Railroad Contract Rates*, Ex Parte No. 358 (Sub-No. 1).

______, *Elimination of Incentive Per Diem Charges*, Ex Parte No. 252 (Sub-No. 5).

______, *Investigation of Adequacy of Railroad Freight Car Ownership, Car Utilization, Distribution Rules, and Practices*, Ex Parte No. 241 (Sub-No. 1).

______, *Liberty Trucking Company, Extension—General Commodities*, 131 M.C.C. 575, 575 (1979).

______, *Master Certificates and Permits*, Ex Parte No. MC-135.

______, *Modification of Rail Carrier General Increase Proceedings*, Ex Parte No. 290 (Sub-No. 3).

______, *Order Granting Railroads Flexibility in Setting Per Diem Rate Levels*, Ex Parte No. 334 (Sub-No. 4).

______, *Policy Statement on Motor Carrier Regulation*, Ex Parte No. MC-121.

______, *Proposal to Repeal Credit Regulations*, Ex Parte No. MC-1.

______, *Rail Market Dominance and Related Considerations*, Ex Parte No. 320 (Sub-No. 1).

______, *Railroad Cost Recovery Procedures*, Ex Parte No. 290 (Sub-No. 2).

______, *Revenue Need Standards in Motor Carrier General Increase Proceedings*, Ex Parte No. MC-128.

______, *Western Railroads-Agreement*, 5B Application No. 2, 364 I.C.C. 1 (1980).

U.S. Court of Appeals for the District of Columbia, *Burlington Northern* v. *United States* (No. 68-2307, D.C. Cir. 1980).

U.S. House of Representatives, Committee on Public Works and Transportation, *Report on Motor Carrier Reform Act of 1980*, Washington, June 1980.

ECONOMICS OF AFFIRMATIVE ACTION

Is Equal Opportunity Enough?

By GLENN C. LOURY*

Affirmative action policies have come increasingly under attack in recent years. Both in the courts and in public discourse questions have been raised about the legitimacy of government efforts on behalf of blacks and other racial minorities.[1] The criticism seems to have two central themes. First, it is argued that those policies which have been tried have not had a noticeable effect on the economic standing of minority group members. (See James Smith and Finis Welch.) They thus constitute yet another example of costly but ineffective government regulation, according to this view. The second theme strikes more deeply at the foundation of these policies. Its adherents argue that even if effective programs could be designed, they ought not be implemented. There have been philosophical and empirical arguments advanced to support this conclusion. Essentially, the philosophical argument states that it is wrong for government to intervene on behalf of certain groups (and thus, necessarily, at the expense of others); this amounts to reverse discrimination—a visiting of the fathers' sins upon the sons.[2] The empirical argument concludes that, moral issues aside, such intervention is unwarranted because the consequences of historical discrimination have been (or will soon be) largely eliminated. (See B. Wattenberg and W. Wilson.)

In this essay I would like to offer a defense of affirmative action policies against the second of these thematic criticisms. That is, I shall hold in abeyance questions concerning the efficacy of particular programmatic efforts, and concentrate instead on whether government should in principle be taking actions to facilitate economic progress for minority group members. This would seem to be the logical first step in constructing an intellectual basis for affirmative action policies. Of course, philosophers and legal scholars interested in theories of distributive justice have devoted considerable attention to this question in the past ten years. (See R. Dworkin and T. Nagel.) The approach adopted here differs from these earlier efforts in two ways. First, I shall endeavor to meet the empirical argument directly, by pointing to evidence which suggests that significant racial economic disparity persists. Secondly, I will treat the philosophical argument in a manner in keeping with the economist's traditional approach to the question of the desirability of *laissez-faire*. This approach is based upon the concept of market failure. Intervention is favored over *laissez-faire* when, because of some externality, the market outcome is inefficient. Below I argue that an analogous "market failure" contributes to the maintenance of economic inequality between racial groups in our society. As such, intervention which redresses this inequality is warranted.

I. The Empirical Argument

Since the passage of civil rights legislation in the early 1960's there have been profound changes in the economic experience of racial minorities in this country. A number of analysts have called attention to this change, observing that traditional discriminatory practices, such as unequal pay to equally skilled workers, have been dramatically reduced. (See Richard Freeman, and Smith and Welch.) Moreover, when the data are disaggregated by cohorts, one finds that the disadvantage in wages of younger minority workers is quite small. (See Smith and

*Professor of economics, University of Michigan.

[1] The arguments of this paper are not intended to apply to affirmative action for women.

[2] This argument is developed at length in N. Glazer.

TABLE 1—PERCENT OF MALE WORKERS UNEMPLOYED, BY RACE AND AGE, 1970–79

Age	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
16–19:White	13.7	15.1	14.2	12.3	13.5	18.3	17.3	15.0	13.5	13.9
Nonwhite	25.0	18.9	29.7	26.9	31.6	35.4	35.1	37.0	34.4	31.5
20+ :White	3.2	4.0	3.6	2.9	3.5	6.2	5.4	4.6	3.7	3.6
Nonwhite	5.6	7.2	6.8	5.7	6.8	11.7	10.6	10.0	8.6	8.4

Source: *Economic Report of the President.*

Welch.) Thus, as younger workers continue to enter the labor force and older workers retire, differences by race in the wages of equally skilled workers may be expected to attenuate. This has led some to question the need for affirmative action, since minority workers seem to be catching up without government help.

This conclusion seems to me premature because it is based on only one aspect of economic status—the earnings of employed workers. No observer of the economic experience of nonwhites in the past decade can have failed to notice that unemployment rates are much higher for minority workers than workers as a whole. Table 1 presents unemployment rates for white and nonwhite male workers by age for the years 1970–79. It is apparent that nonwhite workers are unemployed roughly twice as often as their white counterparts, and that unemployment constitutes a chronic problem for young and nonwhite workers. While these aggregate data do not control for differing individual characteristics (for example, education) which may account for part of this racial disparity, one study of youth unemployment has found that no more than half the racial difference in unemployment rates among young workers can be explained in this way. (See Martin Feldstein and David Ellwood.)

Thus, even if racial differences in the earnings of similarly skilled employed workers were to disappear in the near future, a continuation of current trends in the unemployment experience of nonwhite workers would imply significant economic disparity between the groups. These figures do not prove that minority workers are currently discriminated against in employment opportunities instead of wages, nor do they show that historical discrimination accounts for the currently observed unemployment disparity. However, the data on unemployment certainly suggest that the progress of nonwhite workers in the post-civil rights era cannot be accurately assessed by looking at earnings alone. Moreover, these data are clearly consistent with the hypothesis that there exists racial discrimination in employment opportunities. In a market characterized by excess supply and downward price rigidity (for example, the market for young workers with a minimum wage floor) buyers must use some device for rationing their purchases among the more numerous sellers. The possibility that race is among the characteristics influencing a worker's position in this job queue ought not be ignored.[3]

There is another sense in which comparisons of the annual earnings of racial groups incompletely represent their respective economic positions. The use of cross-section data from a sequence of years does not allow the analyst to discern what happens to the incomes of particular individuals over time. There is some evidence that patterns of year-to-year earnings mobility are quite different for white and nonwhite workers. For example, using longitudinal data, several researchers have found that while the entry level wages of young black and white male workers of similar skills are now quite close, the subsequent rate of wage growth is significantly smaller for the black workers. (See Edward Lazear, Saul Hoffman, and Greg Duncan.) An earlier analysis of occupa-

[3] Charles Betsy has regressed unemployment frequency and duration measures on a variety of explanatory variables, finding quite different coefficients for blacks and whites.

tional mobility among mature male workers showed blacks to be less upwardly mobile out of low-paying occupations and more downwardly mobile out of high-paying occupations than whites (See Bradley Schiller). Moreover, there is evidence that black heads of households in poverty in a given year are considerably more likely than whites to remain in poverty in the following year (see Lee Lillard and Robert Willis), while black families with "high" incomes in one year are much less likely than whites to retain that status in the following year (see my paper with Jerome Culp).

It would appear then that, while the nature of economic inequality between the races has undergone significant change, the gap does not appear to be withering away of its own accord. If this conclusion is accepted, a question then arises as to what, if anything, government should do about it.

II. The Philosophical Argument

This question is a crucial one for advocates of affirmative action. In the period since World War II the principle that an individual's opportunities for advancement ought not depend on race has come to be broadly accepted in our society. Given that racial discrimination in the private sector is not currently practiced, government efforts on behalf of minorities would seem to contradict this basic principle. Advocates of affirmative action now argue that the history of discrimination has created an environment in which equal opportunity alone would not permit minority groups to gain economic parity. Their focus is on the results of the income determination process. Critics, on the other hand, note that one cannot logically urge the necessity of equal treatment while simultaneously demanding special favor. Their focus is on the neutrality of the process itself. This distinction between the fairness of procedures and the fairness of outcomes is a critical one in social philosophy,[4] and in my judgment constitutes the core of the debate over the legitimacy of affirmative action. The nature of the ethical problem should be clear: Racial minorities are undoubtedly worse off today by virtue of the historical use of procedures which did not respect their liberty. Yet, to use the power of the state to "correct" history's wrong doing is to condone disregard for the liberty of those citizens not so favored. The aphorism "two wrongs don't make a right" would seem to apply.

One way to think about this problem is to inquire whether, in theory, we should expect the continued application of racially neutral procedures to lead eventually to an outcome no longer reflective of our history of discrimination. If the answer to this question were negative, then adherence to a policy of equal opportunity alone would condemn those whose rights had historically been violated (and their progeny) to suffer indefinitely from what most would regard as ethically illegitimate acts. Since this would (presumably) be an undesirable state of affairs, a case for intervention would thereby be made. Of course, even if the effects of historical discrimination were to eventually be eroded through the application of racially neutral procedures, this "correction" might take so long as to be of little practical significance. The point here is that there are reasons (to be discussed presently) to believe that our society operates so as to pass on from one generation to the next that racial inequality originally engendered by historical discrimination.

The above discussion is intended to persuade the reader that a certain aspect of the dynamic performance of market economies is important in evaluating the ethical legitimacy of affirmative action. The choice between public policy limited to equal opportunity or extended to affirmative action, I submit, should depend upon the extent to which we are confident of the ability of the market to naturally erode historically generated differences in status between groups. It is in this sense that this choice is analogous to the one economists face when considering whether public intervention in the marketplace is desirable. In the latter instance the ability of *laissez-faire* to attain an efficient allocation of resources is the

[4] See R. Nozick's criticism of "end state" theories of justice. John Rawls adopts an opposing view.

crucial issue. Here I suggest that we focus on the extent to which equal opportunity *eventually* erodes discrimination-induced inequality when judging the appropriateness of intervention via affirmative action.

This question, like the question of when does competition lead to efficient resource allocation, is necessarily a logical query about the operations of an idealized economic system. Like the efficiency question, it may be studied by developing a theoretical model of the social phenomenon at issue, and seeking conditions within the context of that model under which the desired outcome obtains. The basic human capital theory of earnings determination, extended to allow for intergenerational effects, is well suited to an investigation of this sort.[5] Elsewhere (1977) I have pursued this question at some length; however, space limitations necessitate that I merely summarize that investigation here. A model is developed in which job assignments are made under conditions of equal opportunity, based solely on an individual's productive characteristics. However, the individual's acquisition of productive characteristics is favorably influenced by the economic success of the individual's parents. Thus, the deleterious consequences of past discrimination for the racial minority are reflected in the model by the fact that minority young people have less successful parents, on average, and thus less favorable parental influences on their skill acquisition processes. Further, the model posits that families are grouped together into clusters or "communities," and that certain local public goods important to subsequent individual productivity (for example, education) are provided uniformly to young people of the same community. This provides another avenue by which background influences achievement, since the nature of the community to which a family belongs also depends on the economic success of the parent.

In order to pose the question most sharply, it is assumed that all individuals have identical preferences with respect to economic choices, and that an identical distribution of innate aptitudes characterizes each generation of majority and minority workers. Thus, in the absence of historical racial discrimination, we should expect that the economic status of minority and majority group members would be equal, on average. I then inquire whether, in this idealized world, the competitive labor market would function in such a way as to eventually eliminate any initial differences in the average status of the two groups.

The results obtained depend upon whether only income, or both income and race, influence the community to which a family belongs. In the former instance, with some additional reasonable assumptions, one can show that equal opportunity always leads (asymptotically) to equal outcomes. In the latter case, however, it is not generally true that historical differences attenuate in the face of racially neutral procedures. Examples may be constructed in which group inequality persists indefinitely, even though no underlying differences in tastes or ability exist.

This last result arises because, when there is some racial segregation among communities, the intergenerational status transmission mechanism does not work in the same way for minority and majority families. An intragroup externality is exerted, through local public goods provision, by the (relatively more numerous) lower income minority families on higher-income minority families of the same community. Because the racial composition of one's community depends in part on the choices of one's neighbors, this kind of effect cannot be completely avoided by an individual's actions. As a consequence, the ability of equal opportunity to bring about equal results is impaired by the desire of majority (and minority) families to share communities with their own kind.[6] Since this social clustering of the races seems

[5] Such an extension is provided in Gary Becker and Nigel Tomes, and my forthcoming paper.

[6] L. Datcher finds that, for young black males, the racial composition of the community in which they were raised has a significant influence, other things equal, on subsequent earnings. An increase of ten percentage points in the fraction white in the community implied an increase in subsequent annual earnings of 3 percent.

to be a continuing feature of our society, the theoretical analysis leads to the conclusion that intervention may be justified.

III. Conclusion

I have argued that current economic differences between whites and nonwhites are such as to obviate the conclusion that the historical effects of discrimination have (or will soon be) dissipated. Additionally, I have suggested some reasons why a *laissez-faire* policy of equal opportunity, but not affirmative action, could leave minority group members perpetually constrained by historically practiced discrimination. Thus, the second thematic argument against affirmative action, mentioned in the introduction, is deemed unsatisfactory.

REFERENCES

G. Becker and N. Tomes, "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *J. Polit. Econ.*, Dec. 1979, *87*, 1153–89.

C. Betsy, "Difference in Unemployment Experience Between Blacks and Whites," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 192–97.

J. Culp and G. Loury, "In Defense of Public Policy Aimed at Reducing Racial Economic Disparity," *Rev. Black Polit. Econ.*, Summer 1980.

L. Datcher, "Effects of Community and Family Background on Achievement," mimeo, Univ. Michigan 1980.

G. Duncan, "The Determinants of Wage Growth, 1971 – 1976," mimeo., Univ. Michigan 1979.

R. Dworkin, *Taking Rights Seriously*, Duckworth 1977.

M. Feldstein and D. Ellwood, "Teenage Unemployment: What's the Problem," Nat. Bur. Econ. Res. disc. paper, 1979.

R. Freeman, "Changes in the Labor Market for Black Americans," *Brookings Papers*, Washington 1973, *1*, 67–131.

N. Glazer, *Affirmative Discrimination*, Basic Books 1975.

S. Hoffman, "Black-White Life Cycle Earnings Differences and the Vintage Hypothesis," *Amer. Econ. Rev.*, Dec. 1979, *69*, 855–867.

E. Lazaer, "The Narrowing of Black-White Wage Differences is Illusory," *Amer. Econ. Rev.*, Sept. 1979, *69*, 553–64.

L. Lillard and R. Willis, "Dynamic Aspects of Earnings Mobility," *Econometrica*, Sept. 1978, *46*, 985–1012.

G. Loury, "A Dynamic Theory of Racial Income Differences," in P. A. Wallace, ed., *Women Minorities and Employment Discrimination*, Lexington Books, Lexington 1977.

______, "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, forthcoming.

T. Nagel, *Equality and Preferential Treatment*, Princeton Univ. Press, Princeton 1977.

R. Nozick, *Anarchy, State and Utopia*, Oxford 1974.

John Rawls, *A Theory of Justice*, Harvard Press: Cambridge, Mass. 1974.

B. Schiller, "Relative Earnings Mobility in the United States," *Amer. Econ. Rev.*, Dec. 1977, *67*, 926–41.

J. Smith and F. Welch, "Race Differences in Earnings: A Survey and New Evidence," R-2295-NSF, Rand Corp. 1978.

B. Wattenberg, *The Real America*, Doubleday 1974.

W. Wilson, *The Declining Significance of Race*, Chicago 1978.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1980.

Affirmative Action and Its Enforcement

By Finis Welch*

The civil rights movement has navigated from the streets to the courts in a continuing effort to remedy effects of the past and to ensure that reflections of it are increasingly dimmed. From the initial emphasis on publicly provided services (largely schools), public gathering places (buses, restaurants, etc.) and civil rights like voting, the focus has broadened to encompass jobs and the range of issues associated with simple linkages between employment and economic success. There is by now an extensive legal and federal administrative enforcement structure designed to monitor terms of hiring and employment to protect groups that are identifiable on the basis of race, age, ethnicity, religion, or sex. I am sure that there must be a fairly good and simple word that adequately summarizes this structure and its associated machinery, but I cannot find one. The term, *affirmative action*, was first used in this context in an Executive Order which required federal contractors to take affirmative action to eliminate effects of past discrimination and to protect against current discrimination. To many, affirmative action continues to have relatively narrow connotations and to them I apologize because I will use it here in reference to the full apparatus. The alternative is *equal employment opportunity*, but I reject it because I do not think this affirmative action apparatus is designed for equality within groups.

The following list summarizes the most important events of legislation, administrative fiat, and court decisions leading to the antidiscriminatory complex which I call affirmative action.

I. A Chronology of Major Events

The Equal Pay Act of 1963 precluded sex-based discrimination in pay in firms covered by *FLSA* (the Fair Labor Standards Act which includes the federal minimum wage law and overtime provisions, among other things).

Title VII of the Civil Rights Act of 1964. Hands down, this is the most important single piece of legislation involving discrimination in employment. Firms as well as unions and employment agencies were prohibited from discrimination on the basis of race, color, sex, religion, or national origin. Terms of employment were broadly interpreted to include hiring, training, promotion, pay, and termination. Exemptions were extended to governments, educational institutions, and a few others as well as to smaller establishments. The law became effective July 2, 1965. The Equal Employment Opportunity Commission (*EEOC*) was established for monitoring compliance. It processed complaints, served as a conciliator between charging parties and respondents, occasionally recommended cases to the Attorney General, and could file as a "friend of the court" in civil actions. Individual recourse was first in the form of a complaint filed with *EEOC* (and in some cases a prior complaint must have been filed before state and/or local Fair Employment Practice Commissions) and then to the courts.

Executive Order #11246 (*September 1965*) prohibited discrimination on the basis of race, color, religion, and national origin among government contractors (under risk of perpetual debarment from holding contracts) and requires contractors to take affirmative action to ensure a lack of discrimination. The Office of Federal Contract Compliance (*OFCC*) was established for general purpose monitoring and work with *EEO* offices within the federal agencies in support of these regulations.

The Age Discrimination in Employment Act of 1967 prohibited discrimination on the basis of age for workers aged 40 and over but less than 65 years.

Executive Order #11375 (*1967*) amended Order #11246 to proscribe discrimination on the basis of sex.

*University of California-Los Angeles. I am indebted to Iva Maclennan and Walter McManus for assistance; to Ann DuRoss, Lee Lillard, and James Smith for their comments; and to Dale Barone for helping locate data.

In May 1968, *OFCC* established a requirement for federal contractors to present a written affirmative action plan complete with goals and timetables for correcting deficiencies.

In 1971, the Supreme Court in *Griggs vs. Duke Power* (401 U.S. 424) established that requirements for hiring, job placement, or promotion which are facially neutral and possibly neutral in intent which nevertheless have a "disparate impact" on protected groups cannot be used unless they show "demonstrable relationship to successful performance" on the job. This places the burden of proof on employers for requirements like a high school degree or acceptable exam scores to demonstrate that those not satisfying the requirements are unable to perform adequately.

The Equal Employment Opportunity Act of 1972 amended Title VII to add state and local governments and educational institutions as covered employers and to reduce the lower bound on numbers of employees to fifteen in all firms. The Act also gave enforcement powers to *EEOC*, permitting it to bring suit in behalf of charging parties and to issue "right to sue" letters to those sanctioned charges where *EEOC* itself would not carry the suit. Further, it extended coverage to federal employers and gave enforcement responsibility to the Civil Service Commission.

In 1976, the Supreme Court in *Chandler vs. Roudebush* (-U.S.-, 12 FEP 1560) decided that federal employees charging discrimination were entitled to a trial *de novo*. Until this decision some circuits had held that charging employees were only entitled to a review of the administrative procedures used by the Civil Service Commission in judging a claim.

In 1977, in *International Brotherhood of Teamsters vs. United States* (431 U.S. 324, 14 FEP 1514) the Supreme Court held that a bona fide seniority system is lawful even though the employer is guilty of pre-(1964) Act discrimination and even though the seniority system perpetuates effects of pre-Act discrimination.

The Age Discrimination in Employment Act Amendment of 1978 extended the upper limit on "protected" ages to 70 years (except for the federal government where an upper limit was removed altogether). This effectively proscribed mandatory retirement before age 70 (although teachers aged 65–70 with tenure in institutions of higher education were exempted until July 1, 1982).

In October 1978, President Carter consolidated within *OFCCP* the central *OFCC* unit as well as eleven different compliance agencies previously dispersed throughout government. The *OFCCP* has responsibility for Executive Order #11246 as amended. President Carter also indicated that in 1981 consideration would be given to consolidating *EEOC* and *OFCCP*.

In July 1979, responsibility for enforcement of the Age Discrimination Act was given to *EEOC*, which was empowered to sue in claimants' behalf after 1972.

This chronology somewhat belies the lags involved in putting the enforcement machinery in motion. I cannot find statistics for *OFCC* prior to 1968 and in that year, there were only 29 budgeted positions with an appropriation of about $16,000 each. While it is true that *EEO* offices in individual agencies were responsible for monitoring contract compliance it is a reasonable guess that there were no more than six employees in the *EEO* agency offices for each employee in *OFCC*. By 1980 employment in federal contract compliance monitoring had increased seven-fold over 1968 levels. Likewise, *EEOC* started slowly with an initial budget of $2.25 million and 8,700 cases filed in 1966. But its growth has been more rapid and by 1980 the caseload (new filings) was eight times as large as in the initial year and the (constant dollar) budget was twenty-one times as large.

Perhaps most of *EEOC's* growth reflects the fact that until 1972 it had no real enforcement powers. Claims were investigated and conciliation was attempted but the burden of federal court enforcement rested mainly on the charging individual.[1] Since 1972, *EEOC* has had the power to litigate, although I have no information on the magnitude of this activity. Even today, the *EEOC* budget amounts to something less than $1,500 per case "resolved," and one doubts whether the majority are closely examined.

Tables 1 and 2 provide some simple measures of growth in activities of *EEOC* and *OFCC* and in the federal court caseload. While it remains true that the majority of civil rights cases in federal courts are not Title VII cases, it is nevertheless true that

[1] See Phyllis Wallace for a discussion of the early enforcement structure.

TABLE 1—SUMMARY STATISTICS FOR THE TWO ENFORCEMENT AGENCIES TOGETHER WITH CASES FILED IN FEDERAL COURTS PERTAINING TO TITLE VII OF THE CIVIL RIGHTS ACT

	EEOC		Employment Cases Filed in Federal Courts under Title VII	*OFCC*	
Year	Budget ($1,000)	Resolved Cases (1,000)		Budget ($1,000)	Positions
1965	3,250	6.4	[a]	[a]	[a]
1970	13,250	8.5	340	570	34
1975	55,080	62.3	3,930	4,500	201
1978	84,550	80.0	5,500	7,190	216
1979	111,420	81.7	5,480	7,910[b]	216[b]

Sources: *EEOC* budgets are from OMB, The *Budget of the U.S. Government*, various years, *EEOC* cases resolved; *Tenth Annual Report* and *The Budget of the U.S. Government*. Title VII cases filed: *Annual Report of the Director*, various years. The *OFCC* data are from unpublished data.

[a] Not available.

[b] Crude estimates. The 1979 *OFCCP* budget was $44.7 million and reflected consolidation of eleven *EEO* agency offices with *OFCC* to form *OFCCP*. Both the budget and number of positions reported for 1979 are simple extrapolations.

TABLE 2—SUMMARY STATISTICS FOR CIVIL RIGHTS CASES IN FEDERAL COURTS

	Civil Rights Cases				All Civil Cases Tried
		As a Percent of All Cases:			
Year	Numbers Resolved	Resolved	Tried	Percent Title VII (Civil Rights Cases Filed)	Exclusive of those involving Civil Rights
1971	3,970	5.6	8.7	14.6	6,710
1975	7,850	9.2	13.6	37.8	6,950
1979	12,600	10.7	21.3	41.2	7,100

Source Annual Report of the Director, various years.

these cases increased from 8.6 to 41.6 percent of all civil rights cases filed during a decade (1970–79) in which the number of civil rights cases more than tripled. It is interesting that since 1971 there is no significant pattern of growth in the number of civil cases tried in federal courts aside from growth in civil rights cases per se. We do not know the share of Title VII cases among the 21 percent of cases tried that are civil rights, but I assume that it represents something more than two in five.

The scale of employment discrimination litigation is an order of magnitude greater now than a decade ago and one cannot but wonder what the effects have been and what they will become. At this time we have no hard evidence but speculation is fun anyway.

In most cases, evidence of discrimination (henceforth "in employment" is implicit) requires evidence of a general pattern and practice and these cases are inherently statistical. While statistics have been used for several aspects of employment, they are most frequent in studies of hiring and pay. The basic idea is that observed differentials between groups should not exceed what might reasonably have been expected on the basis of chance, although to my knowledge it is perfectly alright for differentials to be too small to have occurred randomly. While you can guess about the variety of machinations that have accompanied claims of statistical proof and their counterparts, the space allotted to me here is insufficient for details.

Briefly then, statistical studies of pay typically use regression to "explain" salaries on the basis of group identifiers and a set of control variables. Just which variables is usually the crux of the issue and seniority is about the only one for which there is a general consensus. Many variables that have commonly been used in professional literature are objected to on the basis of their potential for "disparate impact." Recall that the Griggs decision held that a facially neutral criterion (what could be more neutral than a variable named *X*?) for promotions (and, implicitly, for pay) is seen as impermissible if it has a disparate impact unless it bears a demonstrable relation to successful performance on the job. Irrespective of questions of discrimination between groups, most economists—and I am one of them—would argue that within, say, white males, wages approximate perceived productivity. If so, then the demonstration of a within-group correlation between pay and *X* demonstrates a relationship between productivity and *X*. Clearly a regression using dummy variables for group identification draws all of its information for estimating the effects of *X* from contrasts within groups. But just as clearly, if the inclusion of *X* as an explanatory variable in a pay regression affects the estimate of the between group differential, it must be that *X* has a disparate impact. Thus it seems to me that if the disparate impact criterion is taken literally and if within group pay differentials are not viewed as productivity differentials, then either independent measures of productivity must be used—can you think of any—or we are stuck with comparisons of simple averages or of averages corrected only for seniority.[2]

Alongside criteria for comparing rates of pay, there are questions of hiring which usually consist of contrasts of persons actually hired with what would be expected in random drawings from the "relevant" labor pool. The objective is to ensure proportionate representation of protected minorities or women in the "skill" or highly paid positions. And although, to my knowledge, there is no prohibition against overrepresentation of protected groups in lower-paying positions, I am afraid that there are incentives to avoid doing so. The most obvious is the potential for a charge of "shunting": with proportional minority representation in higher-paying positions and more-than-proportional representation in the others, a randomly drawn minority employee is more likely to be in a lower-paying position than a similar worker taken from the majority or reference group. In all this, I assume that productivity is not observable by enforcement authorities since if it were the equal pay question would be trivial.

Although it is hard to find a simple description for analyzing the constraints a firm faces, I personally do not think much is lost by describing them as a combination of an employment quota and an equal pay constraint. The pay constraint either refers to pay within jobs or job titles or within groups having similar measurable characteristics, *X*. The quota is double-edged even though there is no prohibition on overrepresentation of protected groups in lower-skill or pay positions, a firm having such overrepresentation will on average pay less to protected workers and, therefore, risks charges of unequal pay unless it also has demonstrably valid predictors, *X*, that fully explain the differential.

This is the basis of my concern that vis-à-vis skill acquisition affirmative action serves both as a carrot and a stick for protected workers. For the skill positions, it is a carrot. It creates powerful incentives for employers to increase minority representation in these positions and in attempting to do so they will seek out those whom they consider most productive. But for the unskilled positions, it is a stick which creates contrasting incentives to avoid hiring too many of the seemingly protected workers. I will return to this point with an empirical summary of recent trends in black-white earning differentials, but before then I will outline some of the issues that I consider relevant to evaluating employment quotas. As with most

[2] I won't pursue the question of unobservables except to note that when regressions are used to "correct" crude pay differentials there is the very uneasy question of how within group residuals, which are themselves corrected differentials, are interpreted. Would you argue that all within a group who have negative residuals are discriminated against in favor of those with positive residuals?

social programs, they are issues of efficiency and equity.

Consider a set of employer constraints such that a minimum proportion q of workers must be from a specified group which I will call the minority. There is also an equal pay constraint that holds within jobs or job titles. If we measure ability or productivities in wage units then the combined wage and quota constraints gives a the wage $w=(1-q)a_1+qa_2$, where a_1 refers to the average ability of majority workers within a job and a_2 refers to the average ability of minority workers. The wage as an average of abilities simply reflects the upper limit that profit-maximizing employers can pay. Notice that averaging creates an incentive for assortative matching of employees: since the wage received is an average, majority employees will prefer to be matched with the most productive minority workers and vice versa. The quota constraint is binding at higher levels of ability only if $a_1>a_2$, although as I have twice noted, it will bind throughout the lower ranges irrespective of the $a_1:a_2$ contrasts if enforcement authorities use broadly based averages for salary comparisons.

The question of effects of such quotas on efficiency is pretty much open. There are two dimensions at issue. One involves the implicit minority-majority heterogeneity within job clusters and raises questions whether composite group performance is sensitive to heterogeneity. The other involves choices by workers of occupations etc. Suppose there are two activities, A and B, and that only A is covered by the quota. (If you like, B may refer to nonmarket activities when the quota completely covers the market.) And, suppose that each worker has an ability pair (a, b)—again expressed in wage equivalents—for the respective activities. The endogeneity of these abilities is more relevant to long- than to short-run contrasts, but for the moment, assume that they are exogeneous. Without the quota, activity choice depends only on the $a:b$ comparison. With the quota the comparison is between the wage w, for activity A and the opportunity wage b. Since with $a_1>a_2$ the averaging in w implies a tax on majority workers of $q\ (a_1-a_2)$, and a subsidy to the minority of $(1-q)(a_1-a_2)$, efficiency losses arise from the exit from A by members of the majority whose rent, $a-b$, is less than the tax and the entry into A by members of the minority whose subsidy exceeds their rent $(b-a)$. The efficiency losses consist of the foregone rents.

The efficiency of the quota as a transfer mechanism depends largely on the thickness of the rental margin. If wages are predominately rents, then taxes and their associated transfers have little allocative effect, and to the extent that skills are adaptive, that is, endogenous, it is likely that efficiency in transfers of this type erodes as skills adapt.

I have assumed to this point that the quota is costlessly enforceable, but if the amount of litigation now in process is an indication, we must admit that enforcability is very much an issue. If we presume that affirmative action taxes, like other taxes, are partially avoidable and that the degree of successful avoidance responds to investments, then in contrast to ordinary income taxes there is an asymmetry between taxes paid and receipts available for potential transfer in employment quotas. With a quota, the tax perceived by an employer for hiring a minority worker is a_1-a_2 (the opportunity cost of not hiring a majority worker) and yet the transfer the minority employee receives is at most $(1-q)(a_1-a_2)$, (It is less if the minority employee is marginal to activity A, i.e., if $a_2<b<w$.) Thus, per dollar of potential transfer, we expect employers to invest more in avoidance than we would for each dollar of income tax revenue. But, that is where the transfer disadvantage of the quota stops; for once the tax is paid, the transfer has occcurred. Not so for transfers funded from general revenues: Once the tax is paid, there is the not-so-simple administrative cost of identifying recipients, of maintaining the machinery for the transfer, of administering the transfer (queues for food stamps and the costs of waiting time), and the efficiency costs arising from the incentive effects (on labor supply, marital stability, and possibly on the birth rate) of the transfer itself. To

TABLE 3—BLACK-WHITE EARNINGS RATIOS: MALES, SELECTED YEARS
(Numbers are average earnings of blacks as a percentage of average earnings of whites)

Years of School Completed	1967	1973	1978
I. All Ages			
		A. Annual Earnings	
All	59	65	72
8–11	67	73	74
12	69	74	77
16 or more	63	71	84
		B. Weekly Earnings	
All	62	68	75
8–11	69	75	77
12	73	76	81
16 or more	62	72	88
II. Those Out of School 1-5 Years			
		A. Annual Earnings	
All	69	76	74
8–11	79	77	69
12	81	87	76
16 or more	74	92	98
		B. Weekly Earnings	
All	73	82	81
8–11	85	87	80
12	83	90	85
16 or more	75	92	96

Source: Public Use Tapes from the March *Current Population Surveys*.

Note: The March *Surveys* refer to income (in this case, wages and salaries only) earned last year and to weeks worked last year. The years referenced are income not survey year. Observations are restricted to persons *reporting* income. Persons flagged as having their income imputed (in 1979 the flag is an individual one but in 1968 and 1973 there is only a flag for family earnings) are excluded. Other exclusions include those self-employed or working without pay, those reporting average weekly earnings below $10, and those who reported "major activity last week" as being retired or in school or those who worked less than 50 weeks "last year" and gave retired or school as a reason. Estimates of those in their first five years out of school are inferred from the experience imputation suggested in my paper with William Gould.

my mind, q would have to be very large for the asymmetry between taxes and revenues in quotas to dominate. So long as the only concern is that of a majority-to-minority transfer, the incentive compatabilities between quotas and recipient responses offer attractive alternatives to most of the other mechanisms we have devised.

But, should the concern be only that of equity between groups? Who gains from a quota? It is tempting to argue that the biggest winners are the most skilled who would have fared best anyway, and that is what I expect the data to show. It, however, is not a necessary result of the formal theory where, once the assortment problem is solved, the question is that of the joint distribution between $a_1 - a_2$ and a_2. I do think that a good case can be made for an argument of something akin to a discontinuity where below some minimum ability, minority workers either gain nothing or actually lose, but above that level, it all depends on the relative dispersions of the majority and minority skill distributions.[3]

[3] Where the majority density is disperse relative to the minority density there is a positive association between the transfer and minority ability, and the converse is true when the minority density is relatively disperse.

I would very much like to discuss the demonstration or role modeling effects of affirmative action, but I don't know what to say. I personally hope and suspect that they are important, and if they are, then the incentive compatability noted earlier extends across generations.

In closing, I will summarize my concern about the two edges of affirmative action by reporting simple tabulation from the March annual income surveys of the *Current Population Surveys* (*CPS*). Table 3 provides the summary. My data span every year, 1967 through 1978, but I have selected only the two endpoints and the median year, 1973. The numbers are ratios of average annual earnings and weekly wages for black relative to white men. The data for persons of all ages show a continuing pattern of growth, overall and for selected school completion levels. Yet, the data for recent job market entrants tell a different story. As is true for every cross section I have examined, black-white wage ratios are higher for more recent cohorts, and that pattern seems to hold here when persons of all schooling levels are combined. Yet, patterns between the schooling levels are sharply contrasted. Across the years, there is continuing relative wage growth for college graduates and 1978 shows approximate parity. Yet for the high school dropouts and for those graduating from high school but not going on to college, the period since 1973 exhibits the first persistent decline that I have seen in examining a variety of sample surveys (see my paper with James Smith for a survey of earlier studies) for the periods since 1960.

Clearly the data presented in Tables 1 and 2 suggest that the modern affirmative action push did not gather anything like its present momentum until the mid-1970's. Coincidentally, that is when the relative earnings of less-schooled (and presumably, less-skilled) black men began to fall. It may be pure coincidence and there are many alternatives to explore before further conjecture is warranted. I will note, however, that the continued earnings growth of less-schooled blacks who had entered the job market earlier does not contradict the potentially adverse effects on more recent entrants. The enforcement structure includes provisions for continual examination of reasons for termination and of pay while employed. It does not include provisions for employment of those who already are disproportionately represented.

REFERENCES

W. Gould and F. Welch, "An Experience Imputation or an Imputation Experience," mimeo., Rand Corp. 1976.

B. L. Schlei and P. Grossman, *Employment Discrimination Law*, Washington 1979.

______ **and** ______ *Employment Discrimination Law*: *1979 Supplement*, Washington, 1979.

J. Smith and F. Welch, "Race Differences in Earnings: A Survey and New Evidence," in Peter Mieszkowski and Mahlon Straszheim, eds., *Current Issues in Urban Economics*, Baltimore 1979.

P. Wallace, "Employment Discrmination: Some Policy Considerations," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton 1973.

Administrative Office of the U.S. Courts, *Annual Report of the Director*, Washington, various years.

Equal Employment Opportunity Commission, *Tenth Annual Report*: *A Decade of Equal Employment Opportunity 1965-1975*, Washington 1976.

U.S. Executive Office of the President, *The Budget of the United States Government*, Office of Management and Budget, Washington, various years.

U.S. Office of Business Economics, *Current Population Surveys*, Public Use Tapes.

Monetarist Principles and the Money Stock Growth Rule

By BENNETT T. MCCALLUM*

Given the influence of Milton Friedman, it is hard to keep from identifying "monetarism" with the advocacy of a policy rule that would require the money stock to grow at a constant rate and prohibit cyclical adjustments in government spending or in tax schedules.[1] This identification is somewhat inaccurate since Karl Brunner and Allan Meltzer, the other two leading proponents of monetarism, have not always been advocates of a constant money growth rate. It may nevertheless be useful to relate one's thoughts about monetarism to Friedman's rule, as will be done in this paper. But the question that immediately arises is, what more fundamental beliefs about the economy give rise to the idea that such a rule would be socially desirable. At this more basic level there may be more agreement among monetarists than about the rule itself. In any event, it appears that there are two basic monetarist *propositions* that are of crucial importance, as follows. (*i*) Cyclical and secular movements in nominal income are primarily attributable to movements in the stock of money relative to capacity output. (*ii*) There is no permanent tradeoff between unemployment and inflation or any other characteristic of the path of the price level—that is, the natural rate of unemployment hypothesis is valid. Of course, another belief is essential—that the monetary authority can exert reasonably accurate control over the stock of money if it sets its mind to that task—but it is not a major source of dispute with nonmonetarist economists, so I will not include it on my list.[2]

I. The Natural Rate Hypothesis

Let us begin with a brief explanation of the importance of the second of these propositions, the natural rate hypothesis (*NRH*). To appreciate its significance one needs only to note that, alone, proposition (i) has nothing to say about the division of nominal income fluctuations into price and output components. Thus alone it can have no implications concerning the behavior of an economy's inflation and/or unemployment rates, magnitudes that are of much more inherent importance than nominal income. The crucial nature of proposition (ii) is certainly clear from the writings of Friedman (especially 1968, 1971) and is stressed by Franco Modigliani, but has been neglected in many discussions of the subject.[3]

The other point concerning the *NHR* that needs to be emphasized is that its conditions are, in fact, *not* satisfied by numerous specifications which claim to permit no long-run

*Professor of economics, University of Virginia, and visiting professor, Carnegie-Mellon University. I am indebted to Albert Ando, Robert Barro, Alan Blinder, Karl Brunner, Carl Christ, Robert Hodrick, Bob Lucas, Allan Meltzer, and John Whitaker for helpful suggestions and comments. Also, I am grateful for the financial support of the National Science Foundation (SES 79-15353).

[1] Friedman's rule also requires that the government budget be balanced on average. For compact recent statements of the rule, see Robert Lucas (1980) and Carl Christ (1979, p. 534).

[2] This list has deliberately been kept as short as possible, in order to increase its discriminatory power. Its two items correspond to those stressed by Frank Hahn. The discussion presumes a closed economy.

[3] See, for example, the papers and comments in the volume edited by Jerome Stein. The *NRH* is closely related to the Brunner-Meltzer hypothesis that "the economic system is stable."

tradeoff. As Robert Lucas pointed out in 1972, an expectational Phillips curve—one that relates unemployment to only the unexpected part of the inflation rate—implies the possibility of "*unlimited* real output gains [unemployment reductions] from a well-chosen inflationary policy" (p. 53) unless expectations are taken to be rational. With adaptive expectations, for example, unemployment could be kept low forever if the inflation rate were made to accelerate forever. Of more significance for current debates, given the acceptance since 1972 of the rational expectations hypothesis, is the use of unemployment-inflation relationships not of the expectational variety. A prominent example is provided by the class of formulations that involves the concept of a "nonaccelerating-inflation rate of unemployment," sometimes abbreviated *NAIRU*.[4] If there exists a stable relationship between the rate of unemployment and the acceleration of the inflation rate, as is presumed by the *NAIRU* concept, then evidently there are acceleration magnitudes that will yield a permanently lowered rate of unemployment. Under such formulations, therefore, an argument that rapidly accelerating monetary growth is undesirable needs to be justified by reference to some welfare criterion expressed in more detail than is usual in the monetarist literature. It seems, therefore, that *NAIRU* formulations should be regarded as inconsistent with monetarist doctrine. They are certainly inconsistent with the *NRH* itself.

II. Monetary vs. Fiscal Policy

Having emphasized the importance of proposition (ii), let us now turn to (i) and the sizable body of literature that contrasts monetarism with "fiscalism." A prominent strand of this literature was initiated in the 1973 paper by Alan Blinder and Robert Solow, which emphasized the continuing effects—in a model with wealth terms appearing in *IS* and *LM* functions—brought about by the ongoing issuance or retirement of government bonds required under the money growth policy rule by the government budget restraint. Significant contributions to this strand have been made by Karl Brunner and Allan Meltzer, Carl Christ, Ettore Infante and Jerome Stein, and James Tobin and Willem Buiter. In a later piece, Blinder and Solow (1976) suggested that there are two significant "messages," both of which are very robust to model specification. These are "that *the economy is more likely to be stable if deficits are financed by printing money than if they are financed by floating bonds*" and "that *the long-run effect of government spending on aggregate demand is greater when deficits are bond-financed than when they are money financed*" (1976, pp. 505–06, italics in original). As these messages constitute a direct challenge to monetarist proposition (i), they warrant scrutiny. The following paragraphs will argue that neither message is persuasive. Because the crucial relevant effects have to do with the nature of the government budget restraint, not the workings of the economy, it will be possible to conduct the argument in a setting that is almost model free.

The claim that deficits have larger long-run effects on aggregate demand when bond financed—or, equivalently, that an open-market purchase is contractionary!—is somewhat misleading even at the terminological level. In particular, it refers not to the comparative extent to which equal changes in the stocks of money and bonds cause shifts in the aggregate demand schedule, but rather to system-wide reduced-form effects on nominal income in complete systems that include aggregate supply functions, government budget restraints, income tax schedules, and specifications regarding capital accumulation, as well as aggregate demand (*IS-LM*) relations. In addition, "long run" does not mean after lags in consumption, investment, and portfolio behavior have worked themselves out—there are no such lags in the models in question—but rather that constraints requiring money and bond stocks to be constant are imposed on the system.

Terminology aside, the substantive validity of this second message is highly dubious.

[4] See, for example, James Tobin. The same is of course true for formulations involving a "noninflationary rate of unemployment."

The evident basis for its belief is the following implication of the models of Blinder and Solow (1973), Christ (1979, pp. 533), Infante and Stein (p. 490), and others: unless an implausible condition is satisfied, the system cannot be dynamically stable under the Friedman rule unless bond finance is more expansionary.[5] So a presumption of stability would justify the second message. But reflection upon the meticulous analysis of Christ (1978, 1979) suggests that the correct conclusion is not that the second message is valid, but rather that instability (of certain variables) obtains under the Friedman rule.

To develop an understanding of this result, let us first consider a special model in which it does not quite hold, namely, a purely "Ricardian" model in which the current asset value of government bonds to private agents is precisely offset by agents' recognition of future taxes implied by bond interest payments. As is argued in my earlier paper, the stock of bonds does not, in this case, appear as a variable in any of the behavioral equations of a macro-economic model of the *IS-LM*, expectational Phillips curve variety. Nor do tax variables or parameters. Consequently, the system dichotomizes into two distinct blocks. The first of these includes the model's behavioral relations and explains movements in output Y, the price level P, and the nominal interest rate r, conditional upon time paths of the (high-powered) money stock M and real government purchases G. The second block includes tax-transfer schedules and the government budget identity; it explains (given Y, P, r, M, and G) movements in tax receipts and the nominal stock of bonds, B. If we combine these last relations we can summarize the second block with the following equation, in which T and τ $(0<\tau<1)$ are the intercept and slope parameters in a nominal tax-transfer schedule:

$$(1)\quad DM+DB=GP+rB-T-\tau(YP+rB)$$
$$Dx\equiv dx/dt$$

Using (1), it can easily be seen that the behavior of B is unstable under the monetary rule. In view of the dichotomy described above, (1) can be expressed with $DM=0$ as

$$(2)\qquad DB=(1-\tau)rB+\xi$$

in which ξ is exogenous.[6] Thus, with r exogenous and positive, we have an explosive differential equation in B. The stock of bonds explodes in a positive or negative direction if a deficit or surplus ever occurs.[7]

In the pure Richardian case just considered, the explosion of B has no effect on the variables of primary interest—Y, P, and r. But suppose that capitalization of future tax liabilities is incomplete, so that the real financial wealth of the private sector is $(M+\phi B)/P$, with ϕ "small" but greater than zero. In this *nearly* Ricardian case, which I henceforth assume to be empirically relevant, the system does not dichotomize. Consequently, if B explodes, there will be effects on Y, P, and r that tend to impart explosive behavior to those variables. The question, then, is whether feedback behavior of Y, P, and r into (1) will keep B from exploding.

It seems highly unlikely that such feedback will impart stability to B for the following reasons. With the usual behavioral specifications in which bond wealth $\phi B/P$ is relevant for expenditure (IS) and portfolio (LM) relations, its effects on aggregate demand are positive in the former and negative in the latter. Thus the quantitative importance of B on the instantaneous equilibrium values of Y and P would be relatively small even in the absence of future tax capitalization. And with substantial but incomplete capitalization, as herein presumed, these effects will be substantially reduced. Thus functions relating Y, P, and r to current values of B will involve *weak* relationships; the partial derivatives $\tilde{Y}_1$, $\tilde{P}_1$, and $\tilde{r}_1$ of the reduced-form functions $Y=\tilde{Y}(B, M, \pi)$, $P=\tilde{P}(B, M, \pi)$, and $r=\tilde{r}(B,$

[5] The condition is that the partial derivative of the aggregate demand function with respect to the bond stock be "large" in relation to $1-\tau$, with τ=marginal income tax rate. But this condition is almost certainly invalid, because of substantial tax capitalization.

[6] If DM is a nonzero constant, then the current argument would be expressed in terms of the bond-money ratio, rather than B.

[7] It may be useful to think of r and ξ in (2) as stationary equilibrium values of those variables. Then the result implies local instability.

M, π) will be small in magnitude.[8] Therefore, when these functions are used in (1) to obtain the dynamic representation of B in the near-Ricardian case, as in

$$(3)\quad DB=\left[G-\tau\tilde{Y}(B,M,\pi)\right]\tilde{P}(B,M,\pi)$$
$$+(1-\tau)\tilde{r}(B,M,\pi)B+\text{constant}$$

the implied behavior of B will be well approximated by (2). Thus the behavior of B will be explosive for a wide variety of specifications of the model's behavioral equations.[9]

The foregoing result eliminates, it should be emphasized, the reason at hand for thinking that bond-financed deficits would be more expansionary than money-financed deficits (that open-market purchases would be contractionary). The argument seems to leave us, however, with the conclusion that dynamic instability would prevail under the Friedman rule. It thus seems to support the idea that adoption of the rule would be undesirable.

The instability result has been obtained, however, in a discussion that neglects the effects of economic growth. In a growing economy, equation (1) would continue to hold but the relevant variable for considerations of dynamic stability would not be B, but the *ratio* of bonds to real income, $b=B/Y$.[10] And the appropriate counterpart of (2) would be

$$(2')\qquad Db=\left[(1-\tau)r-DY/Y\right]b+\zeta$$

with ζ analogous to ξ in (2).[11] Furthermore, with money growth as specified by Friedman's rule, it is reasonable to presume that r will be of approximately the same value as the *real* rate of interest. Then, since the latter should be close in magnitude to the (steady state) rate of output growth, it becomes quite likely (with $\tau>0$) that $(1-\tau)r-DY/Y$ will be negative. But that, of course, implies that b will be dynamically stable. Thus, the source of instability provided by the Friedman rule is not present in an economy in which output growth typically proceeds at or above the rate $(1-\tau)r$. The first antimonetarist message, as well as the second, seems then to be unwarranted.

III. Concluding Remarks

It has been argued that the body of literature under discussion provides little reason for skepticism regarding the monetarist proposition (i). Some possibility of instability remains, however, and it must be recognized that while the validity of propositions (i) and (ii) may be necessary, it is not sufficient, for belief in the desirability of a constant money growth rule. It is possible, even if propositions (i) and (ii) are true, that other policy rules would induce operating characteristics of an economy that would be superior to those yielded by constant money growth.

In this regard, it is interesting to recall that in Friedman's original design of "A Monetary and Fiscal Framework for Economic Stability," it was the stock of bonds, not money, that was to be exogenous to cyclical activity. Money stock changes were to be nondiscretionary but were to play the role of financing government deficits and surpluses. Only later, in *A Program for Monetary Stability*, did Friedman propose an exogenous (constant) rate of monetary growth. The original Friedman rule, it should be noted, apparently possesses all the features of automaticity stressed in Friedman's later writings. Furthermore, with money-financed deficits more expansionary than bond-financed, it would imply a strengthening of any stabilizing effects that might be induced by automatic tax responses to fluctuations in nominal income. Its implementation would be difficult, to say

[8] In these functions, π is the expected rate of inflation. It is taken as exogenous in most of the models under discussion; the effect of rational expectations does not seem to alter the results.

[9] In the Brunner-Meltzer system, the result will obtain with bond finance ($\bar{\mu}=0$) since my argument implies that (in their notation, p. 83) $\varepsilon(y,S|0,AM)$ is small.

[10] I am deeply indebted to Albert Ando for calling to my attention the importance of output growth.

[11] To see this, define $m=M/Y$, $g=G/Y$, and write (1) as $Dm+Db+(m+b)DY/Y=Pg+(1-\tau)rb-T/Y-\tau P$. Then impose Friedman's rule by setting $Dm=0$ and requiring g, τ, and T/Y to be constants.

the least, but perhaps the original Friedman rule nevertheless warrants renewed consideration.

REFERENCES

A. S. Blinder and R. M. Solow, "Does Fiscal Policy Matter?," *J. Public Econ.*, Jan./Feb. 1973, *2*, 319–37.

______ **and** ______, "Does Fiscal Policy Still Matter? A Reply," *J. Monet. Econ.*, Nov. 1976, *2*, 501–10.

K. Brunner and A. H. Meltzer, "An Aggregate Theory for a Closed Economy," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976.

C. F. Christ, "Some Dynamic Theory of Macroeconomic Policy Effects on Income and Prices under the Government Budget Restraint," *J. Monet. Econ.*, Jan. 1978, *4*, 45–70.

______, "On Fiscal and Monetary Policies and the Government Budget Restraint," *Amer. Econ. Rev.*, Sept. 1979, *69*, 526–38.

Milton Friedman, "A Monetary and Fiscal Framework for Economic Stability," *Amer. Econ. Rev.*, June 1948, *38*, 245–64.

______, *A Program for Monetary Stability*, New York 1959.

______, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

______, *A Theoretical Framework for Monetary Analysis*, New York 1971.

F. H. Hahn, "Monetarism and Economic Theory," *Economica*, Feb. 1980, *47*, 1–17.

E. F. Infante and J. L. Stein, "Does Fiscal Policy Matter?," *J. Monet. Econ.*, Nov. 1976, *2*, 473–500.

R. E. Lucas, Jr., "Econometric Testing of the Natural Rate Hypothesis," in Otto Eckstein, ed., *The Econometrics of Price Determination Conference*, Washington 1972.

______, "Rules, Discretion, and the Role of the Economic Advisor," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago 1980.

B. T. McCallum, "On Macroeconomic Instability from a Monetarist Policy Rule," *Economic Letters*, 1978, *1*, 121–4.

F. Modigliani, "The Monetarist Controversy or, Should We Forsake Stabilization Policies?," *Amer. Econ. Rev.*, Mar. 1977, *67*, 1–19.

Jerome L. Stein, *Monetarism*, Amsterdam 1976.

J. Tobin, "Stabilization Policy Ten Years After," *Brookings Papers*, Washington 1980, *1*, 19–71.

______ **and W. Buiter,** "Long-Run Effects of Fiscal Monetary Policy on Aggregate Demand," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976.

Monetarist, Keynesian, and New Classical Economics

By JEROME L. STEIN*

Keynesians, monetarists, and new classical economists agree that the steady-state rate of inflation is closely related to the growth of the money supply, and that monetary policy cannot affect the equilibrium rate of unemployment. Disagreement concerns the macrodynamics of unemployment and inflation between steady states. My aim is to state each theory as a refutable set of hypotheses and evaluate their consistency with available evidence.

Keynesian and new classical economics (*NCE*) are polar extremes. Keynesians (James Tobin; Franco Modigliani; Sidney Weintraub) claim that monetary and fiscal policies affect output and employment relatively quickly through their effects upon aggregate demand; but these policies have weak effects upon the rate of inflation. To reduce the rate of inflation significantly requires the creation of Okun Gaps, by any combination of fiscal and monetary policies. Only by adopting an incomes policy can the rate of inflation be reduced significantly without serious iatrogenic (physician induced) effects upon output and employment. To be sure, an incomes policy produces a misallocation of resources but: "It takes a heap of Harberger triangles to fill an Okun Gap" (Tobin, p. 468).

The *NCE* claims (Thomas Sargent; Robert Lucas; Robert Barro) that "... policy makers face no cruel choice between inflation and unemployment over any relevant time frame" (Sargent, pp. 213–14). There is no way that the government can systematically raise or lower the unemployment rate, relative to the equilibrium rate, even in the short run. Many macro-economic models imply that the change in the unemployment rate depends upon lagged values of the unemployment rate, real disturbances and the difference between the current price level and the value which the market anticipated on the basis of information available at any earlier date. A necessary condition for the validity of the *NCE* is the rational expectations hypothesis (*REH*) that the forecast error is a serially uncorrelated term with a zero expectation. It follows that the mathematical expectation of the change in the unemployment rate just depends upon lagged unemployment rates, which reflect frictions in the economy resulting from costs of adjustment. Policymakers cannot systematically change the unemployment rate through monetary policy.

The *NCE* claims that anticipated monetary policy affects the price level quickly and systematically because, in the Quantity Theory equation, it has no systematic effects upon either the level of output or velocity. Inflation can be reduced quickly without mathematically expected iatrogenic effects upon output and unemployment, if the monetary authority were publicly committed to a lower rate of monetary growth.

The monetarist position (Karl Brunner; Alan Meltzer; Milton Friedman) was considered radical a decade ago. At present, it is intermediate between the other two views. Its main tenet is that inflation is primarily a monetary phenomenon. Contrary to the Keynesian view, monetarists claim that a restrictive fiscal policy without a reduction in the rate of monetary expansion cannot reduce the rate of inflation. Although there is no relation between a constant rate of inflation and the unemployment rate, monetarists disagree with the *NCE* and believe that there is a short-run tradeoff between the speed at which inflation is reduced and the temporary rise in the unemployment rate. Even a publicly announced

*Brown University. I am indebted to the following for perceptive criticisms of an earlier draft: W. Bomberger, G. Borts, C. Christ, C. Engel, M. Friedman, M. Goodfriend, W. Haraf, D. Laidler, T. Mayer, B. McGrath, K. Nagatani, W. Poole, R. Solow, and E. Tower.

reduction in the growth of monetary aggregates would lead to a temporary rise in the unemployment rate.

I. Micro-Economic Evidence Concerning the Rational and Asymptotically Rational Expectations

A necessary condition for the validity of the *NCE* is the Muth *REH* that the price anticipated to prevail at time t, given the information available at prior date $\tau, p^*(t;\tau)$, is equal to the subsequently realized price $p(t)$ plus a pure noise term (Muth, equation 5.7). Unlike the implications of adaptive expectations, the forecast error must be a nonsystematic, nonserially correlated, variable. Much recent work in macroeconomics takes the *REH* as an axiom and inquires why the unemployment rate or Okun Gap is serially correlated. I consider the *REH* as an empirical proposition to be tested against an alternative: the asymptotically rational expectations hypothesis *ARE*. The validity of the *REH* cannot be evaluated objectively and directly on a macro-economic level, because there is no objective measure of the anticipated macro-economic price level, or any other macro-economic variable. Survey data concerning macro-economic anticipations are suggestive; but they do not necessarily reflect the anticipations of those who have actually purchased or sold goods, labor services, and financial instruments. Commodity futures, and foreign exchange forward, markets are ideal bases for testing the *REH*, because anticipations are reflected continuously in the markets prices of assets traded in almost perfect markets.

The market price of a futures contract is equal to the price that is anticipated to prevail at its maturity less a positive or negative risk premium (depending on the balance of hedging pressure and risk aversion) plus a random term. Applied to futures markets, the *REH* states that: a regression of the price of a maturing futures contract $q_t(t)$ upon the futures price i months earlier $q_t(t-i)$ should yield a slope which is not significantly different from unity and is significantly different from zero. The intercept term is less significant, since it reflects the average risk premium over the sample period.

I examined 22 futures contracts in wheat, 25 corn contracts, 36 soybeans contracts during the period 1973–78, 9 March soybean contracts 1969–78, and 25 December wheat and 28 December corn contracts 1903–32, at distances $i=1,3,5,7$ months to maturity. The data are closing prices of the contract on the last trading date and closing prices in the middle of the month i months earlier, as reported in the Chicago Board of Trade Statistical Annual. In some cases during 1973–78, there was an overlap of observations. The results are similar for each commodity.

First, there are no significant differences among the means of the prices of the maturing future $q_t(t)$ and the futures prices $q_t(t-i)$ at the various distances from maturity, on the basis of an analysis of variance. The mean futures price i months earlier is an unbiased estimate of the mean price of the subsequently maturing future. Second, the slope of a regression of the price of the maturing future $q_t(t)$ on the futures price i months $q_t(t-i)$ generally decreases monotonically as the distance to maturity increases. For contracts one month distant from maturity, the slope is often not significantly different from unity. On contracts longer than two months to maturity, the slope is either significantly less than unity or not significantly different from zero. For example, during 1974–78 with 25 observations on corn contracts one month prior to maturity the slope is 0.923; and the 90 percent confidence interval is (1.07, 0.78). On corn contracts three months to maturity, the slope is 0.564; and the 90 percent confidence interval is (0.93, 0.19). Third, the correlation between the price of the maturing future and the futures price i months earlier decreases as the distance to maturity increases.

Graphically, for each regression, the mean futures price i months earlier and the mean price of the maturing future lie on the 45 degree line. When $i=1$, the regression line is often not significantly different from the 45 degree line. For contracts more than two months distant from maturity, the regression line rotates through the mean points on the 45 degree line, and has a slope less than unity which approaches zero as the distance to maturity increases.

In the foreign exchange market, the *REH* can be examined by regressing the next period's spot price $p(t+1)/p(t)$ on the forward price of that contract one period earlier $q_{t+1}(t)/p(t)$, both deflated by the current spot price $p(t)$ to eliminate the influence of trend. The *REH* implies that the regression coefficient is not significantly different from unity and significantly different from zero. The intercept term reflects the net hedging pressure and risk aversion over the sample period. Of the nine currencies studied by Richard Levich, (i) seven coefficients are negative; (ii) none is significantly different from zero; (iii) in five cases the 90 percent confidence interval ranges from a number greater than unity to a negative number. These results are inconsistent with the *REH*.

There is the same pattern of forecasting errors in the commodity and foreign exchange markets. In the latter, the forecasting error between the subsequently realized spot price $p(t)$ and the forward price of that contract $q_t(t-i)$ some months earlier has the same sign as the change in the spot price $p(t)-p(t-i)$ between these two dates. David Kaserman observed this in the U.S.-Canadian dollar exchange rate 1955-61; and Levich followed his procedure in his study of nine major currencies 1967–78. In each case, the assumption that the forecast error has no systematic pattern is rejected at the 1 percent level, whether the time span is one month or three months. There is indeed a positive association between the forecasting error and the trend of spot prices: a result which is more in conformity with adaptive than with rational expectations.

In the case of commodities, the same pattern tends to appear. Regress the price of the maturing future $q_t(t)$ on the futures price i months earlier $q_t(t-i)$ and a dummy variable which is $+1$ if the spot price has risen between the two dates and zero otherwise. Two results are noted for maturities greater than one month. First, the regression coefficient of the futures price is brought closer to unity. Second, the sign of the dummy is significantly positive. The forecast error has the same sign as the trend in spot prices.

Based upon the evidence that I examined to date, expectations appear to be asymptotically rational *ARE*. The anticipated price $p^*(t;\tau)$ converges asymptotically to the subsequently realized price $p(t)$ as the time to maturity $t-\tau=i$ decreases. The mean anticipated price is not significantly different from the mean subsequently realized price. However, particularly on contracts more than two months to maturity, anticipations systematically lag behind subsequently realized prices during periods of rising and falling prices, in the manner implied by adaptive expectations. The *REH* is not consistent with this pattern of forecast errors. The logic behind the *ARE* hypothesis is similar to Holbrook Working's (1958) explanation of why an important piece of new information must generate a somewhat gradual price change rather than an instantaneous one.

II. A Macrodynamic Model which can Imply Alternative Hypotheses

My strategy is to consider a dynamic macro-economic model which can imply any of the three views, depending upon the parameter specification. After a terse description of the model (based upon my 1974 model), the statistical hypotheses are explicitly formulated and tested.

The measured unemployment rate $U(t)$ is positively related to the unobserved excess supply of labor; and the latter is positively related to the real wage (adjusted for the level of technical progress). The growth in the real wage is the growth in the nominal wage (which depends upon the deviation of the unemployment rate from its equilibrium value and the anticipated rate of price change) less the rate of inflation. It follows that the change in the unemployment rate $DU(t)$, where $D\equiv d/dt$ depends negatively upon the deviation of the unemployment rate from its equilibrium value U_e, and negatively upon the actual rate of inflation $\pi(t)$ less the anticipated rate of inflation $\pi^*(t)$. Approximate $DX(t)$ by $X(t+1)-X(t)$.

(1) $$DU(t)=U(t+1)-U(t)=$$
$$-a_1[U(t)-U_e]-a_2[\pi(t)-\pi^*(t)]+\epsilon$$

where the coefficients reflect the degree of nominal wage flexibility and the short-run slope of the excess demand for labor equation with respect to the real wage.

The rate of inflation is the sum of the growth of the nominal wage in excess of the rate of technical progress and a function of the Keynesian excess demand gap: desired consumption plus investment plus government purchases less current output. This excess demand gap is the vertical distance between the aggregate demand curve and the 45 degree line; and it is quite different from the negative of the Okun Gap: the horizontal distance between capacity output and current output. On the basis of fairly general consumption and investment equations and an interest rate equation which equilibrates the bond market, the rate of inflation indirectly depends upon: the unemployment rate, real balances $m(t)$, the anticipated rate of inflation, real government purchases g, and the ratio θ of the stock of government interest-bearing debt to money.

(2) $$\pi(t)=P(U(t),m(t),\pi^*(t);g,\theta)$$

The rate of growth of real balances is the growth of the money supply μ less the rate of inflation.

The macro-economic *REH* is that the forecast error in (1), between the anticipated rate of inflation at the time pricing and spending decision are made π^* and the current rate of inflation $\pi(t)$, is a serially uncorrelated term with a zero expectation. It was suggested that this hypothesis seems to be inconsistent with micro-economic evidence.

Consequently, the *NCE* hypothesis is that the mathematical expectation of the change is the unemployment rate from period t to $t+1$, conditional upon the information available at t, just depends upon the history of the unemployment rate and is independent of monetary and fiscal variables at the initial date (see Sargent, pp. 215, 221). The alternative asymptotically rational expectations *ARE* hypothesis is that the change in the anticipated rate of inflation depends upon the difference between the equilibrium rate of inflation implied by the model π_e and the currently anticipated rate of inflation.

(3) $$D\pi^*(t)=c(\pi_e-\pi^*(t)) \qquad ARE$$

There are several noteworthy aspects to the *ARE* equation. The anticipated rate of inflation converges slowly to the steady-state solution π_e as determined by the model and policy inputs. Adaptive expectations do not utilize this information. All three macroeconomic points of view agree upon the steady-state solution, but disagree about the speed of convergence. Insofar as economic agents consist of adherents to the three different points of view, the *ARE* equation captures the areas of agreement and disagreement. The *ARE* equation is consistent with the micro evidence discussed above.

Three sufficient conditions generate monetarist results. (a) Inflationary anticipations are generated by the *ARE* hypothesis. (b) The position along a given aggregate demand curve supply does not affect the rate of inflation significantly: $P_1 \approx 0$ in equation (2). (c) Bond-financed fiscal policy has two countervailing effects. A rise in government purchases raises aggregate demand directly; but the rise in the ratio of bonds to money, generated by the budget deficit, raises interest rates which reduces demand: $P_5<0$ in equation (2). On balance, the effect upon excess aggregate demand is substantially less than would occur if government spending were financed by money. Over a period of a year, bond-financed fiscal policy has a weak effect upon aggregate demand.

These conditions imply the following monetarist scenario. A rise in the rate of monetary expansion initially raises real balances, since prices change differentially. The Keynesian excess demand for goods is raised and, regardless of the current value of the Okun Gap, the rate of inflation is increased. The asymptotically rational anticipated rate of inflation is raised slowly because economic agents have different models concerning the dynamics of inflation between steady states and there is uncertainty whether the monetary change is transitory or permanent. The increase in the anticipated rate of inflation raises the inflation of unit labor costs and aggravates the effect of the rise in the Keynesian excess demand for goods. The net effect is that the actual rate of inflation rises faster than the anticipated rate. The decline in real unit labor costs reduces the unemployment rate. Eventually, the actual and anticipated rates of inflation catch up to the constant growth of the money supply. When that occurs, if the unemployment rate deviates from its equilibrium value, nominal wages grow at a different rate than prices. The adjustment in real unit labor costs restores the unemployment rate to its equilibrium value.

Mathematically, the change in the rate of inflation from period t to $t+1$ is a function of the change in real balances from $t-1$ to t. In discrete time, the rate of inflation from t to $t+1$ is a weighted average of the rate of monetary expansion $\mu(t)$ and rate of inflation $\pi(t)$ from $t-1$ to t. This monetarist equation implies that monetary policy can change the rate of inflation without first producing a series of Okun Gaps. It does so by changing the rate at which the aggregate demand curve is rising vertically.

Since the unanticipated inflation is positively related to the growth of real balances, the change in the unemployment rate equation (1), given the monetarist parameter specification, is

$$(4)\quad U(t+1)-U(t)=-a_{11}[U(t)-U_e] \\ -a_{12}[\mu(t)-\pi(t)]+\epsilon$$

Not only does the change in the unemployment rate from t to $t+1$ depend upon the initial unemployment rate but, contrary to the *NCE*, it also depends upon the difference between the rate of monetary expansion less the rate of inflation from $t-1$ to t.

III. An Evaluation of the Evidence

The *NCE* hypothesis is evaluated relative to monetarist hypothesis (4), concerning the change in the unemployment rate or Okun Gap, using annual observations from 1958–77. The *NCE* hypothesis is that the expectation of the change in the unemployment rate from $t-1$ to t, conditional upon the information available at $t-1$ is fully described by an autoregressive equation (see Sargent, p. 215). The *ex post* disturbance, which the *NCE* often associate with "unanticipated money growth," is unknown at $t-1$ and its expectation is zero. Hence, it does not feature in the tests considered here, which are exclusively based upon information known no later than $t-1$. Monetarist hypothesis (4), or regression (6), states that, since prices change differentially and expectations are asymptotically rational, the expectation of the change in the unemployment rate from year $\mathrm{t}-1$ to t, given the information available at $\mathrm{t}-1$, depends upon the unemployment rate in year $\mathrm{t}-1$ and the growth in real balances from year $\mathrm{t}-2$ to $\mathrm{t}-1$.

$$\begin{array}{llll} (5) & U(t)-U(t-1)=1.77 & -0.28U(t-1) & \\ & (t=) \qquad\qquad (1.14) & (-1.134) & \\ & -0.19U(t-2) & +0.18U(t-3) & NCE; \\ & (-0.633) & (0.581) & \\ & & R^2=0.21;\ SE=1.126 & \end{array}$$

$$\begin{array}{lll} (6) & U(t)-U(t-1)=3.8 & -0.663U(t-1) \\ & (t=) \qquad\qquad (5.34) & (-5.19) \end{array}$$

$$\begin{array}{l} -0.403[\mu(t-1)-\pi(t-1)]\ \mathit{MONETARIST}; \\ (-5.47) \end{array}$$

$$R^2=\ 0.701;\ SE=0.668;\ DW=1.8$$

The monetarist hypothesis demonstrably outperforms the *NCE* hypothesis, on the basis of information known at $t-1$. The R^2 is raised from 0.21 to 0.70 and the standard error is reduced from 1.126 to 0.668. Each coefficient in the Monetarist equation is significant at the 1 percent level. It is most unlikely that the *NCE* and monetarist hypotheses are observationally equivalent. On average over the period, the implied equilibrium rate of unemployment is 5.73 percent.

A fundamental difference between the Keynesians and the monetarists concerns the link between changes in the rate of monetary expansion and changes in the rate of inflation. Keynesians believe that the change in the rate of monetary expansion must first affect the unemployment rate; and the resulting change in the Okun Gap leads to a change in the inflation. Since the Okun Gap could be changed as well by appropriate fiscal policy, there need be no systematic relation between monetary policy and the rate of inflation. Monetarists believe that the change in the rate of inflation depends upon the change in the Keynesian excess demand which depends upon the change in real balances, regardless of the value of the Okun Gap. However, as equation (6) states, the change in real balances also effects the subsequent unemployment rate. There is indeed a social cost involved in reducing the rate of inflation; but it is considerably less than that claimed by the Keynesian and considerably more than that claimed by the *NCE* (see L. Meyer and R. Rasche).

A direct test of the competing Keynesian and monetarist hypotheses is described by (7). The dependent variable is the change in the rate of inflation of the *GNP* deflator $\pi(t)-\pi(t-1)$ from year $t-1$ to t, using annual data 1958–77. The independent variables are the initial year's unemployment rate $U(t-1)$, and the change in real balances $\mu(t-1)-\pi(t-1)$ from year $t-2$ to $t-1$. The monetarist hypothesis is that the significant variable is the change in real balances, whereas the Keynesian hypothesis is that it is the Okun Gap or unemployment rate.

$$(7)\quad \begin{aligned}\pi(t)-\pi(t-1) &= \underset{(0.607)}{0.792} \; \underset{(-0.63)}{-0.148U(t-1)} \\ (t=)\quad & \quad + \underset{(3.172)}{0.425[\mu(t-1)-\pi(t-1)]};\end{aligned}$$

$$R^2=0.465; \quad SE=1.224$$

In each case, the monetarist hypothesis is consistent with the data and the Keynesian hypothesis is rejected. A direct regression of the rate of inflation π(t) on the previous year's rate of inflation π(t−1) and rate of monetary expansion yields equation (8).

$$(8)\quad \begin{aligned}\pi(t) &= \underset{(-0.717)}{-0.436} + \underset{(4.38)}{0.569\pi(t-1)} \\ (t=)\quad & \quad + \underset{(3.81)}{0.545\mu(t-1)};\end{aligned}$$

$$R^2=0.802; \quad SE=1.21; \quad DW=2.28$$

The inflation rate for Canada 1961–78, and for the world as a whole 1953–79, are remarkably similar to the *U.S.* equation. These inflation equations support the monetarist view that inflation is primarily a monetary phenomenon whose driving force is the previous year's growth in real balances.

"The beauty and power of a theory are dependent on its ability to capture with clarity and simplicity the key elements of a complicated process." The *ARE* hypothesis and monetarist model seem to dominate the Keynesian and *NCE* in that respect.

REFERENCES

D. Kaserman, "The Forward Exchange Rate," *American Statistical Association*, Bus. and Econ. Statistics, 1973, 417–22.

R. Levich, "Further Results on the Efficiency of Markets for Foreign Exchange," Fed. Res. Bank, Boston 1978.

L. Meyer and R. Rasche, "On the Costs and Benefits of Anti-Inflation Policies," *Fed. Reserve Bank St. Louis Rev.*, Feb. 1980, *62*, 3–14.

J. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

T. Sargent, "A Classical Macroeconometric Model," *J. Polit. Econ.*, 1976, *21*, 207–37.

J. L. Stein, "Unemployment, Inflation and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867–87.

J. Tobin, "How Dead is Keynes?," *Econ. Inquiry*, Oct. 1977, *15*, 459–88.

H. Working, "A Theory of Anticipatory Prices," *Amer. Econ. Rev. Proc.*, May 1958, *48*, 188–99.

Stabilization, Accommodation, and Monetary Rules

By JOHN B. TAYLOR*

A central feature of the monetarist approach to the problem of inflation is a preannounced gradual reduction in monetary growth. This reduction is to be sustained until a monetary growth consistent with a zero, or an acceptably low, target rate of inflation is reached. Thereafter, monetary growth is to be held constant at this new rate. The specifics of this prescription differ by the type of monetary measure used for calculating growth rates—either high-powered money, a monetary aggregate, or nominal *GNP*—and by the length of the transition period during which the noninflationary growth is approached.[1] The prescriptions are alike in ruling out contingent deviations from the plan should economic conditions change, either during the transition period or after the beginning of the constant growth rate rule. The plans are rigid, having no explicit contingencies.

This rigidity has been the target of most critiques of monetarism. Stressing the inefficiencies which a noncontingent monetarist rule would entail, many economists have pointed out the advantages of alternative rules which react to economic events in a structured and stable way. Stanley Fischer and J. Phillip Cooper showed, through a series of examples, that these inefficiencies of monetarist rules exist even when lags are long and variable.

The present paper is an attempt to examine quantitatively which operating characteristics of a monetarist rule are inefficient, and which are relatively efficient. Certain, if not all, features of monetarism could operate efficiently within the context of a particular model and a certain set of parameter values. Few econometric studies, however, have been designed to estimate the specific differences between a monetarist rule and an efficient rule, and to determine whether these differences are statistically significant. Is the monetarist rule inefficient because it is not countercyclical, or is it inefficient because it does not accommodate inflation? Or is it inefficient on both counts?

The particular model used for this analysis is adopted directly from my 1979 econometric investigation. It is a small model with rational expectations and certain rudimentary types of inflation inertia. The model fits the *U.S.* economy fairly well, and some new complex variable techniques developed and applied to the model by David Livesey, enable a simple analytic treatment of efficient policy choice in place of less transparent numerical analysis. Using the results of Livesey, the differences between monetarist and efficient rules can be clearly and rigorously shown without reliance on numerical optimization or simulation techniques.

The results indicate that the monetarist nonaccommodation of inflation seems to work reasonably well and, although some value judgements are required, does not generate significant output instabilities. On the whole, however, the monetarist rule is inefficient. The inefficiencies arise mainly because of its rigidity with respect to business cycle developments. Some moderate countercyclical monetary responses can help to stabilize the economy. In fact, a classic countercyclical monetary policy combined with no accommodation of inflation is nearly efficient.

*Princeton University. I am grateful to Albert Ando, Phillip Cagan, Edmund Phelps, David Livesey, and Jerome Stein for helpful discussions, and to the National Science Foundation for research support.

[1] Allan Meltzer concentrates on the monetary base, while at the other extreme, William Fellner concentrates on general monetary-fiscal aggregate demand management and the growth of nominal *GNP*. Milton Friedman and Robert Lucas focus on monetary aggregates for the constant growth rate rule. Lucas presents a more general treatment of gradualism than that described here, and Phillip Cagan discusses gradualism in terms of the degree of economic slack.

I. A Macro-Economic Model

The model assumes that there is a natural or average rate of unemployment which is insensitive to aggregate demand policy, and that the role of monetary policy is to stabilize fluctuations in unemployment around this average rate. Relative to some views of policy activism, this framework already assumes away a certain role for aggregate demand policy—that of sustaining unemployment rates below this natural rate through the use of stimulative policy.[2] According to the model, such low rates could only be sustained with constantly accelerating inflation (at best) and therefore are not consistent with any reasonable notion of price stability. Of course, the remaining role for monetary policy is by no means unimportant or trivial in a quantitative sense, even in relative terms. Another feature of the model, which a few years ago would have made it seem monetarist from the start, is that the main instrument of aggregate demand management is assumed to be the money supply, not fiscal policy or explicit interest rate policy.

Algebraically the model can be represented in the following equations:

$$(1)\quad y=\beta_1 y_{-1}+\beta_2 y_{-2}+\beta_3(m-p) +\beta_4(m_{-1}-p_1)+\beta_5\hat{\pi}+\eta+\theta_1\varepsilon_{-1}$$

$$(2)\quad \pi=\pi_{-1}+\gamma\hat{y}+\varepsilon+\theta_2\varepsilon_{-1}$$

where y is the percentage (positive) deviation of real output from potential output, m is the *log* of the money supply, p is the *log* of the price level, and π is the rate of inflation. The hats represent rational forecasts, and η and ε are serially uncorrelated shocks to aggregate demand and inflation, respectively. A theoretical rationale for these equations is provided in my earlier paper. The first equation represents the impact of money on aggregate demand, and the second equation represents the impact of expected aggregate demand on inflation. In my earlier paper I tested, along with other things, whether the constraints that money does not appear directly in equation (2) are satisfied and found that it was difficult to reject that hypothesis. It is important to note that the measure of potential output used in (1) must be such that when the economy is operating at that level there is no tendency for inflation to accelerate according to (2). According to my previous calculations, this property is satisfied by a potential output measure which is uniformly about 3 percent below the measure of potential output published in the 1980 Annual Report of the Council of Economic Advisers. Quarterly data for the United States during the period 1954–75 yield the following parameter values for the model: the β parameters are 1.17, $-.32$, .58, $-.48$, $-.45$, and $\gamma=.02$, $\theta_1=.38$, $\theta_2=-.67$.

The shocks to the equations are important for our analysis of monetary accommodation. According to most interpretations of the price shock and accommodation issue (see Edward Gramlich, for example), there are certain shocks to the price level (energy, agriculture, a wage push passed through to prices, etc.) which are temporary in that they do not get incorporated in the underlying rate of inflation. The argument for accommodation rests on the temporary nature of these shocks: because they are short-lived they should be matched point-for-point by an increase in the money supply to prevent a decline in aggregate demand. The main problem with this argument is that it is extremely difficult to determine in practice whether such a shock will get incorporated in the inflation rate or not. In reality we observe a mixture of shocks, some which are temporary and some which are permanent. The case for accommodation then rests on playing the statistical averages, partially accommodating by a proportion equal to the fraction of shocks which are temporary

[2] It is possible within this framework to interpret this average unemployment rate as related to the average (target) inflation rate. This relationship would simply be another consideration in determining these targets. One reason for such a relationship might be that stabilization of aggregate demand requires fluctuations in the real rate of interest below zero which is not possible with a zero rate of inflation. A target inflation rate of 3 percent rather than zero would permit such fluctuations and thereby raise the average rate of unemployment.

on average. Of course, eventually it becomes evident whether a shock gets incorporated in the rate of inflation or not. If it does, then the accommodation issue has a different form: it involves a question of accommodating—presumably at most partially—the underlying rate of inflation.

These kinds of considerations enter the present model through the moving average error term in equation (2). A well-known time-series result is that a mixture of a temporary shock (serially uncorrelated) and a permanent shock (random walk) results in a moving average relationship like that shown in (2). If θ_2 is negative, then it represents the proportion of a mixed shock which on average is temporary.

A neglected part of most discussions of accommodation is whether these temporary price shocks have any direct influence on aggregate demand. In terms of the behavior of real balances, it seems reasonable to expect that money demand would respond differently to a price which was known with some probability to be temporary. While measured real balances would surely fall due to the price shock, the demand for measured real balances might also fall. To capture these possible influences the lagged price shock is added to equation (1).

II. Efficient Policies

As stated earlier the objective of policy in this model is to stabilize the fluctuations of output around potential output. The instrument of stabilization policy is *m*. However, according to (2), inflation is influenced by the behavior of output as well, so stabilization efforts must balance two goals—inflation and output stability—according to the weights in the objective function. The stabilization problem gives rise to a tradeoff between output and price stability.

The efficient stabilization policy rule given these objectives has the form

$$(3) \quad m-m_{-1}=h_1y_{-1}+h_2(y_{-1}-y_{-2}) \\ +h_3(m_{-1}-p_{-1})+h_4\pi_{-1}+h_5\varepsilon_{-1}$$

The first two terms represent what is usually called proportional and derivative control: if h_1 and h_2 are negative, high output levels and speedups in the business cycle call for lower monetary growth. The h_3 parameter is a corrective factor for the level of real balances and acts much like an integral control. The accommodation parameters are h_4 and h_5. If h_4 is positive, then there is at least partial accommodation of inflation inertia, while h_5 determines the accommodation of inflation surprises. Note that the h_3 parameter also involves some accommodation because the lagged price p_{-1} enters through this term.

Using complex variable techniques, Livesey has derived a set of convenient analytic expressions for the *h* parameters. These are given by

$$(4) \quad h_1=-(\beta_2+\beta_1)/\beta_3$$

$$(5) \quad h_2=\beta_2/\beta_3$$

$$(6) \quad h_3=-1-\beta_4/\beta_3$$

$$(7) \quad h_5=-\theta_2(h_4-1)+\theta_1/\beta_3$$

The objective of this paper is to determine under what conditions h_1 through h_5 will equal zero, for this defines the monetarist rule. Equations (4) through (6) uniquely determine h_1 through h_3 given the parameters of the model. No value judgement is required to determine these parameters. The parameter h_4 and hence h_5 depend on the weights of inflation fluctuations and output fluctuations in the social welfare function, however, and therefore require a value judgement.

III. Accommodation or Countercyclical Stabilization

Examining first the countercyclical parameters h_1 and h_2, it is clear that these parameters are large and negative (−1.81 and −.55, respectively) when evaluated at the estimated structural parameter of the model. Even with considerable damping of the stabilization policy to reflect parameter uncertainty as described by William Brainard, these parameters would not be set to zero as with a monetarist rule. According to these estimates, a countercyclical stabilization policy whereby monetary growth increases when the economy is in a recession, but not so fast if the recovery appears to be proceeding too rapidly, would significantly reduce business cycle fluctuations.

To judge the appropriate values for the accommodation parameters h_4 and h_5, the variance of output and inflation associated with these parameters must be evaluated. When $h_4 = 0$ (the monetarist value and the nonaccommodative value for the efficient rule) the variance of output is not at its minimum. By raising h_4 above 0 and at the same time adjusting h_5 according to (7), policy becomes more accommodative and this reduces the size of the business cycle swings in output as one would expect. However, the quantitative reduction in output fluctuations is relatively small. The standard deviation of these output fluctuations is reduced by about .18 percentage points while the standard deviation of inflation fluctuations is increased by about 1.5 percentage points. At the value of $h_4 = 0$, the tradeoff is rather unfavorable if one is interested in reducing output swings any further. It is not unreasonable to suggest that $h_4 = 0$ is near the socially preferred point. Viewed geometrically with output stability on the vertical axis and inflation stability on the horizontal axis, the tradeoff between these two goals is very flat at this point. According to (6), when h_4 is near zero, we also have that h_5 is near zero (specifically, if $h_4 = 0$ then $h_5 = -.02$). That is, if there is no accommodation of inflation inertia, it is optimal not to accommodate inflation surprises either. This result depends heavily on the estimated value of θ_1 which measures the extent to which the negative impact of a price shock on aggregate demand is attenuated if the price shock is temporary, and on the estimated value of θ_2 which measures the inflation durability of a price shock. Both of these values are large enough that no special accommodation of price shocks (in the specific sense described here) is required.

The final parameter to consider is h_3. The size of this parameter depends on whether the level or the change in money balances dominates in the monetary impact on the deviations of output from potential. If $\beta_4 = -\beta_3$, then only the change matters and $h_3 = 0$. If $\beta_4 = 0$, then only the level matters and $h_3 = -1$. For the estimated structural values the change effect dominates ($\beta_4 / \beta_3 = -.83$), so that h_3 is small ($-.17$), though not as negligible as h_4 and h_5.

To summarize, the efficient rule is unlike the monetarist rule in its countercyclical reaction to the state of the economy (h_1 and h_2 are far from the zero values of the monetarist rule), but surprisingly similar to the monetarist rule in not accommodating inflation (h_3, h_4 and h_5 are relatively close to zero). It is in this sense that a nonaccommodative countercyclical policy would work well and might be close to efficient. It would perform better than a monetarist rule which is nonaccommodative but which is also noncountercyclical.

IV. Transitional Problems

Little has been said in this technical analysis about the transition problem of moving from one policy rule to another. Clearly the countercyclical policy suggested here in which money growth is permitted to deviate from its long-run target (say 3 percent per year for $M1$-B) only when the economy drifts away from potential, is not the policy which has been in operation on average over the last ten or fifteen years in the United States. Hence a transition problem arises if one is interested in implementing this countercyclical policy.

Although very little research has been done on the subject, one suspects that transitions will be smoother if the policy operating *rules* change slowly. It is the policy rule which in part determines the economic environment in which individual expectations and decisions are made. A rule is technically defined as a set of feedback parameters which describe how monetary policy operates. In this paper a rule is defined by the five h parameters. A small change in a rule corresponds to a small change in these parameters. If a smooth transition requires slow change in rules, in this sense, then the gradualist monetarism prescription described in the introduction to this paper will not result in a smooth transition. Gradualism involves a quick switch from an apparently very accommodative policy rule to one with no accommodation (a change from h_4 near 1 to h_4 near zero in the notation used above). Perhaps a smoother way to proceed would be to shrink h_4 slowly. Such an alternative transition method could be applied to the rule suggested here as well as to the monetarist rule. Practical implementation of this transition method is as difficult to describe as implementation of policy rules in general. Econometricians estimating policy rules during the next ten years may find that h_4 is shrinking, but will probably find it difficult to pinpoint a sharp break. If so, then the transition will be about as smooth as one could expect.

V. Concluding Remarks

The aim of this paper has been to point out some of the good and bad characteristics of monetarist rules. In a particular sense, and for a particular model, I showed that monetarism scored relatively high on the accommodation issue, but low on countercyclical stabilization issues. The low score may not be surprising given the nature of this model, but the high score does seem surprising. The obvious policy implication is to suggest rules which score high on both issues. These would react to business cycle developments, but would not accommodate inflation. Further research to determine whether these results are robust to plausible modifications of the model would be useful. One modification which is the subject of my current research would place more emphasis on anticipatory wage determination by distinguishing between contracts and expectations as the source of the inflation inertia.

REFERENCES

W. C. Brainard, "Uncertainty and the Effectiveness of Policy," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 411–25.

Phillip Cagan, *Persistent Inflation*, New York 1979.

William J. Fellner, "The Core of the Controversy about Reducing Inflation: An Introductory Analysis," in his *Contemporary Economic Problems*, Washington 1978.

S. Fischer and J. P. Cooper, "Stabilization Policy and Lags," *J. Polit. Econ.*, July/Aug. 1973, *81*, 847–77.

Milton Friedman, *A Program for Monetary Stability*, New York 1959.

E. M. Gramlich, "Macro Policy Responses to Price Shocks," *Brookings Papers*, Washington 1979, *1*, 125–66.

D. A. Livesey, "Stabilization Policy: A View from the Complex Plane," CARESS working paper no. 80-09, Univ. Pennsylvania 1980.

R. E. Lucas, "Rules, Discretion, and the Role of the Economic Advisor," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago 1980.

A. H. Meltzer, "The Case for Gradualism in Policies to Reduce Inflation," in *Stabilization Policies: Lessons from the 1970s and Implications for the 1980s*, Federal Reserve Bank of St. Louis, 1980.

J. B. Taylor, "Estimation and Control of a Macroeconomic Model with Rational Expectations," *Econometrica*, Sept. 1979, *47*, 1267–86.

What is Left of the Multiplier Accelerator?

By OLIVIER J. BLANCHARD*

One of the few undisputed facts in macroeconomics is that output is hump shaped, or more precisely that the distribution of weights of the moving average representation of the deviation of quarterly output from an exponential trend has a hump shape. The first eight weights of the distribution are given in Table 1, column 1. Nearly equivalently, output is well characterized by the following $AR(2)$:

$$Y_t = \underset{(16.4)}{1.34}\ Y_{t-1} - \underset{(-5.21)}{.42}\ Y_{t-2} + \varepsilon_t$$

Y: logarithm of real quarterly GNP minus linear trend; sample period: 47–3 to 78–4.

This implies that, following a movement of output from its equilibrium value this period, we expect a movement of output further away from equilibrium for three more quarters before it returns to equilibrium. It also implies that, given only the past history of output, we can predict—some—turning points from expansion to recession (i.e., sequences $EY_{t+i} > EY_{t+i-1}$; $EY_{t+i} > EY_{t+i+1}$) or the reverse; this would not be the case if the best representation of output was a first-order autoregressive process for example.

I. The Multiplier Accelerator

The traditional explanation of the hump shape relies on the dynamics of *private spending* and the combination of the multiplier and accelerator mechanisms.

In its original form (see Paul Samuelson), it is given by

$$C_t = \alpha Y_{t-1}; \quad I_t = \gamma(Y_{t-1} - Y_{t-2})$$

$$Y_t = C_t + I_t + G_t \Rightarrow$$

$$Y_t = (\alpha + \gamma) Y_{t-1} - \gamma Y_{t-2} + G_t$$

where G is autonomous expenditures. This has the required implication for $(\alpha+\gamma) > 1$. In this case white noise disturbances in G generate a hump shape for output.

The large macro-econometric models also generate a hump-shaped response of output, although not to white noise but to serially correlated disturbances in G. This is shown in Table 1, columns (2) to (4), for the MPS model. Although interest rates and prices are endogenous, the hump shaped response of output comes from the IS dynamics: interest rates and price movements only dampen the effect of G. These IS dynamics in turn are explicitly constructed around the multiplier accelerator mechanism (see Carol Corrado for the MPS model).

If we consider column (3), it is characterized by substantial anticipated movements in private spending. Although the initial movement in period 1 is in response to the unanticipated shock in G and therefore unanticipated itself, the movements in period 2 and following are anticipated. It is the existence of such anticipated movements in either consumption or investment which has recently come under attack. It has been argued that, *given interest rates* and tax rates, most movements in consumption and investment are due to new information and that there cannot be large anticipated movements in consumption or investment. If the argument is correct, the multiplier accelerator and, with it, the IS dynamics of large macro-econometric models are misleading and we should look elsewhere for an explanation of the hump shape.

The rest of the paper reviews the theoretical arguments. The next sections present first the case for the prosecution and then for the defense.

*Harvard University. I thank Andy Abel, Ben Friedman, Francesco Giavazzi, Robert Shiller, and Larry Summers for discussions, the National Science Foundation and the Sloan Foundation for financial assistance.

TABLE 1—RESPONSE OF REAL *GNP*

Quarters	*Y* (1)	*G* (2)	*A* (3)	*Y* (4)
1	1.00	1.00	.16	1.16
2	1.32	.90	.43	1.33
3	1.47	.81	.64	1.45
4	1.50	.73	.82	1.55
5	1.30	.65	.76	1.41
6	.99	.59	.72	1.31
7	.80	.53	.64	1.17
8	.79	.47	.36	.83

Note: Col. (1): Estimated response of *GNP* to a one-time deviation of 1 of its own disturbance in quarter 1. (*MA* representation derived from an *AR*(12) estimated on 1947–3 to 1978–4.) Cols. (2), (3), (4): Simulated response of *GNP*(*Y*) and private spending (*A*) to a one-time deviation of 1 in ε, with $G=.9G_{-1}+\varepsilon$(*G*: government expenditures) in the *MPS* model.

II. An Epidemic of Martingales

For *consumption*, the point was made by Robert Hall. Assuming that an individual maximizes expected lifetime utility, he will always act so as to equalize current marginal utility and discounted expected future marginal utility. Formally, if C is consumption, δ the discount rate, r the interest rate assumed known and constant for our purposes, and Ω_t his information set at time t:

$$(1)\quad U'(C_t)=\frac{1+r}{1+\delta}E[U'(C_{t+1})|\Omega_t]$$

Under the further assumption that utility is quadratic (or else as an approximation), this gives

$$E\left[\frac{1+r}{1+\delta}C_{t+1}-C_t|\Omega_t\right]=0$$

The implication is that even if δ is different from r, anticipated movements in income, as they belong to Ω_t, do not lead to anticipated movements in consumption. Furthermore, if $\delta=r$, there are no anticipated movements in consumption: consumption follows a martingale. Otherwise, if δ is different from r, it follows a sub- or a supermartingale.

As a sum of (sub, super) martingales is a (sub, super) martingale, aggregate consumption, excluding those who enter and leave the consumption pool each period, is also a (sub, super) martingale. Even if individual discount rates differ both from the interest rate and across individuals, the proposition that anticipated movements in income do not lead to anticipated movements in consumption remains valid. When the new and disappearing consumers are taken into account, this proposition and the martingale characterization are only approximations.

For *investment*, the point was made in connection with the "q theory" of investment. This theory assumes internal costs of adjustment for capital and derives a relation between investment and the ratio of the market value of the firm to its replacement cost. It has the following structure: Assuming, for example, risk-neutral shareholders, a constant interest rate r, a depreciation rate δ, and quadratic costs of adjustment, value maximization implies the following behavior:

$$(2)\quad \frac{I_t}{K_t}=\delta+\phi(q_t-1)$$

$$(3)\quad q_t=\sum_{j=0}^{\infty}\left(\frac{1+r}{1-\delta}\right)^{-j}E(MR_{t+j}|\Omega_t)$$

where MR_{t+j} is the marginal revenue from a unit of capital in the firm at time $t+j$. (The exact characterization depends on particular assumptions such as whether capital installed is instantaneously productive and so forth. A closely related derivation is given by Andrew Abel.) The characterization is intuitive: the rate of investment is a linear function of a shadow price q; this shadow price is the present discounted value of expected marginal returns to capital.

Being a present value, q_t satisfies the following relation which is implied by (3):

$$(4)\quad E\left(q_{t+1}-\left(\frac{1+r}{1-\delta}\right)(q_t-MR_t)|\Omega_t\right)=0$$

This is a familiar relation for asset prices, usually referred to as an arbitrage or no excess return relation. A similar relation with dividends instead of MR_t holds for stock prices for example. (Indeed, under further assumptions, such as a *CRTS* technology, competitive factor and product markets, q is equal to the price of a share, i.e., the title to one unit of capital as valued in the stock market.)

Equation (4) implies that $X_{t+1} \equiv q_{t+1} - ((1+r)/(1-\delta))(q_t - MR_t)$ is a fair game with respect to Ω_t but not that q_t itself follows a martingale. It has been suggested however that q_t follows approximately a martingale. If this is the case, equation (2) implies that the rate of investment follows also approximately a martingale.

To be sure, the above theories have fairly restrictive—and mutually inconsistent (risk-averse consumers but risk-neutral shareholders for example)—assumptions. The intuition behind equations (1) to (4) suggest however that more complex specifications, such as better treatments of uncertainty, are unlikely to yield drastically different implications. They therefore suggest that given interest rates, consumption and investment movements are likely to be mostly unanticipated.

III. Anticipated Movements in Investment

Intuition suggests that the martingale "approximation" is simply wrong for investment: If costs of adjustment are nearly linear, the firm will adjust its capital stock mainly to current demand conditions; if movements in demand are partly anticipated and partly unanticipated, we would expect both large anticipated and unanticipated movements in investment. If, on the other hand, adjustment costs are very convex, the firm will change its capital smoothly. We would then expect both small anticipated and unanticipated movements in investment. In neither case would we expect the ratio of anticipated movements to unanticipated movements to be necessarily small, as required by the martingale approximation.

To see this more clearly, we can solve equations (2) and (3) for two different values of the convexity parameter. The coefficient ϕ in (2) is directly related to this parameter: the more convex adjustment costs are, the lower ϕ. A value of $\phi = .05$ implies that a rate of net investment of 5 percent annually entails a marginal installation cost of 100 percent of purchase price per unit installed. This may reasonably be taken as an upper bound on the convexity of adjustment costs. A value of $\phi = .5$ implies that a rate of net investment of 5 percent entails a marginal cost of 10 percent per unit. This may be taken as a lower bound. Further assume in both cases that the production function is Cobb Douglas, with a share of capital of 25 percent, the depreciation rate is 12 percent annually, the real interest rate is 3 percent, that firms take output as given and that the wage always equals the marginal product of labor. These heroic assumptions allow us to derive the following equation for annual net investment, IN_t:

$$(\phi = .5) \;\; IN_t =$$

$$.20\left[.503\sum_{0}^{\infty}(.69)^i E(Y_{t+i}|\Omega_t) - K_{t-1}\right]$$

$$(\phi = .05) \;\; IN_t =$$

$$.04\left[.290\sum_{0}^{\infty}(.82)^i E(Y_{t+i}|\Omega_t) - K_{t-1}\right]$$

The first term in brackets can be thought of as the desired capital stock. Higher adjustment costs ($\phi = .05$) imply that more weight is given to the distant expected future. They also imply a slower adjustment to the desired stock. If we further assume for example that output is equal to a constant plus white noise ε_t, we get

$$(\phi = .5) \;\; IN_t = .80IN_{t-1} + .1(\varepsilon_t - \varepsilon_{t-1})$$

$$(\phi = .05) \;\; IN_t = .96IN_{t-1} + .01(\varepsilon_t - \varepsilon_{t-1})$$

Higher adjustment costs lead to higher serial correlation, i.e., smaller anticipated movements but also to smaller unanticipated movements. To summarize, investment does not follow a martingale. For plausible values of ϕ ($\phi = .5$ for example), the traditional accelerator theory—with a modified definition of the desired capital stock—still holds and there can be substantial anticipated movements in investment.

What is the empirical evidence on ϕ? Equations such as (2) have recently been estimated and their results are puzzling: they yield implausibly low values of ϕ, usually around .05. Such values of ϕ imply, as we have seen, very large adjustment costs and very small anticipated or unanticipated movements in investment; this is hard to reconcile with the actual movements of investment. There are reasons to believe that these estimates of ϕ are biased downwards: the shadow price q is usually approximated in the regressions by the ratio of market value to replacement cost which is likely to be a mediocre proxy. The market value itself is also surprisingly volatile during the sample period. (This is a puzzle in itself; see Robert Shiller, 1981.) Both reasons would lead to a downwards bias in $\hat{\phi}$.

If investment does not follow a martingale, where did the martingale "approximation" argument of the previous section go wrong? It went wrong in assuming that the fair game property of X_t implies that the present discounted value q_t follows, even approximately, a martingale. Present discounted values do not in general follow martingales: the present discounted value of an $AR(1)$ variable for example follows also an $AR(1)$ with the same coefficient of serial correlation. This was emphasized by Shiller (1979, Appendix A), but the mistake is still quite frequently made.

IV. Anticipated Movements in Consumption

The story is different for consumption. Under the life cycle hypothesis and the additional assumptions made in Section I, individual consumption indeed follows a martingale. Is it true however of aggregate consumption? Can we really disregard the effects of the change in the consumption pool? Suppose, for example, that aggregate income is deterministic and grows at rate γ. Suppose also that $r=\delta$ so that the consumption of each individual is constant. If we look at the total consumption of the agents present in two successive periods, it is constant. Aggregate consumption however grows at rate γ: disregarding the change in the consumption pool is clearly not innocuous. If aggregate income has also a stochastic component, then to the extent that the expected consumption of those who start consuming is different from the consumption of those who stop, there will be some anticipated change in consumption.

To see whether this anticipated change can be large, consider a world in which one agent is born each period and lives for N periods, in which $r=\delta=0$ and in which aggregate income follows $Y_t=a+\rho Y_{t-1}+\varepsilon_t$, each agent receiving $1/N$ of aggregate income. We can derive the behavior of aggregate consumption C_t and look at the ratio of the anticipated change in C_t to its total change by computing

$$A_N=\frac{E(E(C_{t+1}|\Omega_t)-C_t)^2}{E(C_{t+1}-C_t)^2}$$

For $\rho=1$, i.e., if income itself follows a martingale, then $A_N=0$: aggregate consumption also follows a martingale. If on the other hand, $\rho=0$, then A_N is given by

$$A_N=\sum_1^N\left(\frac{1}{n}\right)^2\Bigg/\left[\sum_1^N\left(\frac{1}{n}\right)^2+\left(\sum_1^N\frac{1}{n}\right)^2\right]$$

As the sum in the numerator converges but the second sum in the denominator diverges, A_N tends to zero as N gets large. This implies that the martingale approximation is correct for large N. If we assume for example that agents consume for 50 years, we get $A_{50}=7$ percent and the martingale approximation is quite good. To reject the martingale result, we therefore have to reject some of the assumptions made in Section I.

An obvious candidate is the implicit assumption that wealth can be negative. What happens if wealth cannot be negative for an individual, if there are liquidity constraints? Consider an individual for whom $\delta=r$, so that in the absence of liquidity constraints, his consumption follows a martingale. How will he plan consumption if he expects to be liquidity constrained? He will still never anticipate to decrease his consumption: if he did, there would be a path of constant consumption involving saving now and dissaving later which would yield higher utility and be feasible. He may, however, anticipate to increase his consumption if he anticipates his income to increase: he cannot borrow against future income. Liquidity constraints therefore imply that consumption may not follow a martingale but do not imply that anticipated decreases in income lead to anticipated movements in consumption.

Recent empirical evidence suggests the existence of liquidity constraints. Marjorie Flavin finds that the effect of current income on consumption is too strong to be consistent with the life cycle hypothesis. Robert Hall and Frederic Mishkin, using micro data, conclude that liquidity constraints may affect approximately 20 percent of consumers. This suggests that, although liquidity constraints exist, they may not be so prevalent so as to lead to large anticipated increases in consumption in response to increases in income.

V. Conclusion

Boldly stated, the conclusions of this paper are that anticipated movements in output—especially anticipated decreases—will not lead to changes in consumption but may lead to large changes in investment. In this sense, the multiplier is dead and the accelerator alive.

Given these conclusions, can we, for example, generate a hump-shaped output only from the dynamics of private spending in response to disturbances in autonomous spending? In response to a disturbance, consumption will react and adjust to a new constant level; the anticipated movements in private spending must therefore come from investment.

If disturbances are unanticipated, their occurrence will lead to an increase in consumption and an increase in investment. Over time, consumption will remain approximately constant and investment will decline as the capital stock adjusts. The overall response of output will not be hump shaped, but declining over time.

If disturbances are anticipated, however, say two quarters in advance, they will lead to an initial jump in consumption, an anticipatory increase in investment. After they have occurred, investment will slowly decline. Overall, anticipated white noise disturbances can generate a hump shape of output. Whether the hump shape of output is actually explained by such anticipated disturbances is a different question that this paper does not address, much less answer.

REFERENCES

A. Abel, "Investment and the Value of Capital," New York: Garland Publishing, Inc., 1979.

C. Corrado, "The Steady State and Stability Properties of the MPS Model," unpublished doctoral thesis, Univ. Pennsylvania 1976.

M. Flavin, "The Adjustment of Consumption to Changing Expectations about Future Income," *J. Polit. Econ.*, forthcoming.

R. Hall, "Stochastic Implications of the Life Cycle Permanent Income Hypothesis: Theory and Evidence," *J. Polit. Econ.*, Dec. 1978, *86*, 971–87.

______ **and F. Mishkin,** "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," mimeo., July 1980.

P. Samuelson, "Interactions Between the Multiplier Analysis and the Principle of Acceleration," *Rev. Econ. Statist.*, May 1939, *21*, 75–78.

R. Shiller, "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *Amer. Econ. Rev.*, forthcoming.

______, "The Volatility of Long Term Interest Rates and Expectations Models of the Term Structure," *J. Polit. Econ.*, Dec. 1979, *87*, 1190–229.

Bankruptcy, Liquidity, and Recession

By BEN S. BERNANKE*

This paper examines the possibility that the economy-wide level of bankruptcy risk plays a structural role in the propagation of recessions. The argument is as follows: Bankruptcy imposes net social costs, so that all agents have an interest in avoiding it. Consumers and firms do this by being careful to retain sufficient liquid assets to meet fixed expenses; banks and other lenders, by being selective in choosing borrowers and limiting the size of loans. The onset of recession strains the system by reducing the flow of income available to meet current obligations and by increasing uncertainty about future liquidity needs. There is a general attempt to insure solvency which leads to a reduced demand for consumer and producer durables—which may in turn generate further income reductions.

I find that this story helps explain some features of recession. It is largely but not entirely supported by the available evidence.

I. Bankruptcy

By many measures, the general level of bankruptcy risk is a (roughly coincident) countercyclical variable. The link from recession to bankruptcy is not principally reduction in net worth, but increased incidence of technical insolvency—the inability to meet current cash obligations. Recessions create financial distress by narrowing the margin between cash flow and debt service.

The financial difficulties induced by recession are not socially costless. Administrative and legal expenses in bankruptcy are substantial. Other costs include losses from hasty liquidation of assets; delays and uncertainties in payments; loss of customer and credit relationships; and interrupted production.

Why, if there is a social gain from avoiding them, do bankruptcies ever occur? Existing explanations rely on some sort of missing market argument (see Jeremy Bulow and John Shoven, and references therein). We can suggest a solution based on a moral hazard: Lenders cannot observe the objective conditions on which borrowers base their portfolio decisions. If a lender does not develop a reputation for pressing his claims, borrowers have an incentive to become too illiquid in order to force an improvement in terms.

Because bankruptcy has net social costs, borrowers and lenders adopt strategies to reduce bankruptcy risk. We are interested in the implications of these strategies for the business cycle. In what follows, assume that there has already occurred an unanticipated drop in national income—because of a drop in government demand, for example. A key result is that, when bankruptcy considerations are taken into account, this drop in income may substantially reduce expenditure on illiquid, durable assets.

II. Consumers

The individual forming a consumption plan is usually thought of as facing a stock constraint—that lifetime consumption not exceed lifetime resources. This restriction can be more generally formulated as a flow constraint—that at each moment there must be sufficient cash and new credit to cover expenditure and debt repayment. The usually ignored flow constraint reduces to the stock constraint only if all assets are liquid (easily convertible into cash), or if the consumer can borrow freely against lifetime wealth. Neither condition holds in practice. Most consumer assets have some degree of

*Stanford University, Graduate School of Business. The section on firms draws heavily on the work of Bruce Greenwald who interested me in this subject. Research assistance was provided by Jon Zucker and John Caskey. The National Science Foundation furnished support through a grant administered by the National Bureau of Economic Research, and grant no. DAR-8007779.

illiquidity—notably human capital, but also durables, some part of equity in business or property, pensions, insurance, and some financial assets. Many assets are likewise imperfect as collateral, because of moral hazard (as in the case of human capital), difficulties in transferring property rights (human capital, pensions, expected bequests), or differences in valuation between borrower and lender due to asymmetric information (durables).

When the flow constraint is relevant, a principal effect of a drop in current income is the reduction of expenditure of illiquid, long-lived assets (such as durables). There are two reasons for this. First, lower current income increases the short-run probability that the flow constraint will have to be satisfied through costly means, for example, the distress sales of assets, borrowing at unfavorable terms, severe reduction in current living standard, or, as the last resort, bankruptcy. As Frederic Mishkin (1976) first pointed out, deferring purchases of durables lessens this danger by conserving financial resources and avoiding additional fixed obligations (debt service). These benefits are felt at once, while the costs are spread over the future. Thus, for nonmarginal falls in income, this strategy will likely be preferred over responses which involve largely current costs, for example, increased labor supply or smaller consumption of perishables.

Second, a drop in current income typically has ambiguous implications for the consumer's estimates of future income flows and, hence, for the level of durables holdings consistent with maintenance of solvency in the long run. An asymmetry arises here: Because durables are illiquid, it is more costly to correct (what turns out to be) an overpurchase than an underpurchase. Assuming that waiting for new information will tend to resolve the ambiguity created by the initial income fall (for example, was the drop caused by a permanent demand shift or by temporary cyclical conditions?), even a risk-neutral consumer will be motivated to defer durables purchases until the uncertainty is resolved. (For a parallel analysis of this "consumer confidence" effect in the context of irreversible investment theory, see my earlier paper.)

III. Firms

With modifications, this story for consumers can be applied to firms. The firm, too, must reconcile long-term spending plans with the necessity of having the cash flow to meet short-term obligations. Low internal liquidity and many fixed expenses increase the risk of financial embarrassment; at least, they raise the cost of new financing. Again, deferral of capital expenditures is an appropriate defense of the balance sheet against a fall in current income.

A consideration that weakens the analogy from consumers to firms is that, unlike consumers, firms have the option of equity finance. It is interesting to note, however, that debt-equity ratios are countercyclical, a fact due only in part to slowdown in internal equity buildup during recession (see Merton Miller). Managers apparently prefer to see equity from a position of relative financial strength. A conjecture is that they are concerned about the signal they send to financial markets. Offering a new issue with an unfavorable balance sheet is like a cut in the dividend; it conveys pessimism on the part of insiders.

The argument for a link between asset composition and spending by consumers and firms is related to the older notion of "liquidity constraints" (formalized in modern treatments of "effective demand"). That essentially static idea, that an agent might literally reach zero liquidity and thus be prevented from making planned purchases, was used (for example) to justify explicitly temporary tax cuts as an expansionary tool. This did not seem credible, since only a small fraction of the economy is at zero liquidity at a given time. Our dynamic formulation recognizes that, while few are actually at zero liquidity or going bankrupt, the need to guard against the risks of low liquidity is a factor in the economic behavior of most agents.

IV. Banks

The basic proposition, that consumers and firms react to a cash squeeze by cutting durables expenditures and avoiding new debt, could be overturned if lenders' terms became sufficiently attractive in recession.[1] Interest rates do tend to be procyclical; one wonders why cycles aren't smoothed out by marginal adjustments in the timing of durables purchases.

A case can be made that the "easiness" of credit in recession is illusory, as follows: Banks are not interested in the extending of credit *per se* but in the probability of repayment. If the risk of default were known and exogenous in each case, the market for loans could be cleared by an interest rate that included a premium compensating for that risk. However, if borrowers have better knowledge of their own default risk than the bank has, then the familiar adverse selection mechanism makes the quality of the applicant pool (unobservable by banks) dependent on the loan terms offered. As a promising line of research beginning with Dwight Jaffee and Thomas Russell has shown, under these circumstances the loan market is cleared not just by interest rates, but by other dimensions of the transaction, such as loan size. A sensible strategy for the bank may well be to set a low interest rate and to limit loan sizes below what borrowers would like. Thus interest rates are only a partial measure of credit tightness.

With the increase in insolvency that accompanies a drop in national income, banks suffer a deterioration in portfolio quality. They become cautious and try to shift toward safer assets. The predicted effects (which are consistent with the data) are a fall in safe asset yields; an increased spread between safe yields and personal or corporate borrowing rates; and reduced willingness of banks to lend to consumers and the (lower-liquidity) firms most likely to require external financing for projects.

The opportunities for arbitrage profit from borrowing to buy durables in recession are thus more limited than they may first appear. The caution of banks, the disadvantages of incurring more debt with a weakened balance sheet, and the temporary uncertainties associated with a fall in income all act to push the marginal cost of debt finance in recession above the posted interest rate. What is likely to make these factors decisive is that the relevant margin is not between buying now and never buying the durable, but between buying now and possibly buying later.

V. Implications for Recession Dynamics and Policy

Costly bankruptcy and asset illiquidity (arising from the absence of certain Arrow-Debreu markets) bring a distinctly Keynesian flavor to the analysis of recession. Expenditure once more depends on income as well as wealth, because of the relationship of current income to liquidity and insolvency risk. Thus the hypothesized initial fall in income may be propagated through the economy via a multiplier mechanism: income affecting spending affecting income. Note that the largest impact is predicted to occur in sectors producing long-lived, illiquid goods; this is in fact what we observe. This description of recession dynamics does not rest on the assumption of sticky prices (except for the fixity of nominal money and debt values). As lower incomes force down real demands, the movement of firms back along their supply curves will cause the general price level to fall (relative to trend). The lower price level will be expansionary through the familiar Pigou and Keynes effects; but in the present model it will also exert a depressing influence by increasing the real (inside) debt burdens and insolvency risks of consumers and firms. If the latter effect is strong enough, the process may be unstable:[2] this was recognized fifty years ago by Irving Fisher.

If the path based on local price adjustment process is unstable, then there is a

[1] It could also be overturned if the supply of durable goods were inelastic over the cycle. However, the large variance of production in these industries suggests a relatively elastic supply in the short run.

[2] This instability argument can be made formally in a flexible-price temporary equilibrium model if one assumes costly bankruptcy, imperfect liquidity of nonmoney assets, and the existence of noncontingent (nominal) debt. In this context, a rise in the price of money (deflation) may perversely increase the demand for money (and reduce the demand for goods) as debtors try to maintain solvency in face of an increased real debt burden. This assumes an equilibrium exists, which is problematic in a model with bankruptcy.

potential justification for public action (macro-economic policy). Tax relief financed by bonds, for example, can be thought of as the government becoming a financial intermediary that makes it possible for low-liquidity agents to borrow. Monetary policy can exploit the nonneutrality inherent in the fixity of nominal debt values to increase spending through "reflection." Of course, this rationalization for policy has no answer to the argument that activist policy is too imprecise for practical use.

VI. Evidence

At present the empirical evidence relevant to the story I have told is limited and does not permit firm conclusions. Econometric results drawn from time-series data support the idea that balance sheet variables help determine durables expenditure. Otto Eckstein, Edward Green, and Allen Sinai reported favorable results in a regression explaining auto purchases; the Data Resources model today makes use of liquidity and debt variables in predicting both consumer and firm spending behavior. Mishkin has done the most extensive work by far (see his 1977 paper and references therein), consistently reporting positive effects of liquid assets and negative effects of debt on various components of consumer durables expenditure and housing. It has been argued, however (see discussion of Mishkin, 1977), that some of these results may derive from spurious correlations; for example, both the stock market (the most variable part of liquidity) and durables purchases could be driven by a third variable, say, expectations of economic growth.

The small amount of cross-sectional evidence offers less support. James Tobin, in an example to illustrate his dichotomous-dependent-variable procedure, found that the ratio of car purchases to disposable income was related to age but not to holdings of liquid assets. Tobin's data base was the Survey of Consumer Finances for 1952–53; I have done some preliminary work using the same survey for 1969. I have found auto expenditures to be best explained by (previous year) disposable income, pre-existing auto stocks, and demographic variables—notably age. The dominance of disposable income over measures of net wealth supports the liquidity approach; contrary to expectations, however, holdings of nonstock financial assets, stocks, and home equity all failed to contribute to the explanation of auto expenditures. The three components of consumer debt included in the regression entered negatively, but at low statistical significance levels.

For firms, there is a well-known tradition that the level of internal liquidity is important for the rate of investment (see, for example, Robert Coen), but I am not aware of a full-fledged balance-sheet analysis of firm investment.

Besides econometric analysis, evidence can be gathered from case study. Mishkin (1978, 1979) has looked at consumer durables expenditure in the Great Depression and the Great Recession of 1974–75. I have made some effort to expand on his work and to look at other episodes; initial results suggest that additional research would be worthwhile. Following are just a few comments.

The Great Depression. Bankruptcy risk was, of course, very important in 1929–33, a period in which banks as well as borrowers had to hoard liquidity in order to maintain solvency. Mishkin has already documented how the liquidity/bankruptcy model can explain the "autonomous" drop in consumer spending to which Peter Temin attributed part of the depression's severity. Other points which can be clarified by this approach include 1) the behavior of interest rates in 1931–32, when (safe) government bill yields and (risky) corporate bond yields radically diverged (Temin, p. 104); 2) the lack of response of investment spending to apparently cheap money later in the depression (not a liquidity trap, but rather a dearth of solvent private borrowers); and 3) the inability of deflation (which increased debt burdens and insolvency) to stimulate the economy.

The 1968 tax surcharge. The failure of this explicitly temporary tax measure, in apparent contrast to the permanent tax cut of 1964, convinced many macro economists of the empirical dominance of pure net wealth (life cycle) theories of consumption. My reported regression results notwithstanding, this conclusion is probably too hasty. While the impact of the surcharge was not trivial ($36 billion in 1969), powerful offsets included easy money and credit policies, a bull market, and (because of the 1966–67 credit crunch) a low inherited debt burden. Further research into this episode is planned.

The 1980 recession. The surprising strength of consumer spending in 1979 (attributed by some to the positive effects of the real estate boom on consumer liquidity) was overmatched by the unexpectedly restrictive credit policies which began in October. Interest rates have followed a familiar path since then: high at first, then falling. Some large firms have had well-publicized cash flow problems. Bankruptcy, default, and debt burdens have approached 1975 highs while durables demand has been sharply cut. The evidence is admittedly impressionistic and circumstantial. Still, the roles of bankruptcy and liquidity factors in this as well as other episodes bear investigation.

REFERENCES

B. Bernanke, "Irreversibility, Uncertainty, and Cyclical Investment," Nat. Bur. Econ. Res. work. paper no. 502, July 1980.

J. I. Bulow and J. B. Shoven, "The Bankruptcy Decision," *Bell J. Econ.*, Autumn 1978, *9*, 437–56.

R. Coen, "The Effect of Cash Flow in the Speed of Adjustment," in Gary Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971.

O. Eckstein, E. Green, and A. Sinai, "The Data Resources Model: Uses, Structure, and Analysis of the U.S. Economy," *Int. Econ. Rev.*, Oct. 1974, *15*, 595–615.

I. Fisher, "The Debt-Deflation Theory of Great Depressions," *Econometrica*, Oct. 1933, *1*, 337–57.

D. Jaffee and T. Russell, "Imperfect Information, Uncertainty, and Credit Rationing," *Quart. J. Econ.*, Nov. 1976, *90*, 651–66.

M. Miller, "Debt and Taxes," *J. Finance*, May 1977, *32*, 261–75.

F. Mishkin, "The Household Balance Sheet and the Great Depression," *J. Econ. History*, Dec. 1978, *38*, 918–37.

______, "Illiquidity, Consumer Durable Expenditures, and Monetary Policy," *Amer. Econ. Rev.*, Sept. 1976, *66*, 642–54.

______, "What Depressed the Consumer? The Household Balance Sheet and the 1973–75 Recession," *Brookings Papers*, Washington 1977, *1*, 123–64.

Peter Temin, *Did Monetary Forces Cause the Great Depression*?, New York 1976.

J. Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, Jan. 1958, *26*, 24–36.

Estimated Effects of the October 1979 Change in Monetary Policy on the 1980 Economy

By RAY C. FAIR*

On October 6, 1979, the Federal Reserve announced what most people interpreted as a change in monetary policy. The purpose of this paper is to estimate the effects of this change on the 1980–81 economy. The effects of the change are estimated from simulations with my model of the *U.S.* economy (1976,1980b). One of the equations in this model, which is discussed in detail in my 1978b paper, is an equation explaining the behavior of the Federal Reserve. In this equation the Fed is estimated to "lean against the wind," i.e., to allow short-term interest rates to rise (fall) in response to an increase (decrease) in real economic activity, in the rate of inflation, and in the past growth rate of the money supply. The change in monetary policy is estimated by adding three dummy variables to this equation: one each for 1979IV, 1980I, and 1980II. The estimated coefficients of these variables are taken to be the estimated effects of the monetary policy change on short-term interest rates.

To estimate the effects of the policy change on the economy, two dynamic simulations were run for the 1979IV–1981IV period: a "base" run that included the Fed behavioral equation with the dummy variables, and a second run that included the equation without the dummy variables. The difference between the predicted values from these two runs for each endogenous variable and each quarter is an estimate of the effects of the policy change on the variable in the quarter. Standard errors of the effects have also been estimated, and these are presented below. The standard errors were estimated by means of a stochastic simulation procedure that I have recently proposed (1980a).

I. The Equation Explaining Fed Behavior

The equation explaining Fed behavior, estimated for the 1954I–1980II period by two-stage least squares, is

$$(1)\quad r_t = \underset{(3.97)}{-13.4} + \underset{(16.77)}{0.874}\, r_{t-1} + \underset{(1.87)}{0.0512}\,\%PD_{t-1} + \underset{(3.96)}{0.0421}\, J^*_t + \underset{(2.16)}{0.0557}\,\%GNPR_t + \underset{(1.52)}{0.0188}\,\%GNPR_{t-1} + \underset{(2.47)}{0.0324}\,\%M_{t-1} + \underset{(3.35)}{1.58}\, D794_t + \underset{(3.02)}{1.59}\, D801_t - \underset{(3.78)}{2.22}\, D802_t$$

$$\hat{\rho} = \underset{(2.16)}{0.246}, \quad SE = 0.444,$$

$$R^2 = 0.965, \quad DW = 1.82$$

where r is the three-month Treasury bill rate, $\%PD$ is the percentage change at an annual rate in the price deflator for domestic sales, J^* is a measure of labor market tightness, $\%GNPR$ is the percentage change at an annual rate in real GNP, $\%M_1$ is the percentage change at an annual rate in the money supply, and *D794*, *D801*, and *D802* are dummy variables that take on a value of one in the relevant quarter (1979IV, 1980I, and 1980II, respectively) and zero otherwise. $\hat{\rho}$ is the estimate of the first-order serial correlation coefficient. The t-statistics in absolute value are in parentheses. A description of the data and the precise definitions of the variables are contained in my 1976 book and 1980b article.

Equation (1) states that the current bill rate is a positive function of the lagged rate

*Yale University.

of inflation, of the current degree of labor market tightness, of the current and lagged rates of growth of real *GNP*, and of the lagged rate of growth of the money supply. Lagged values of these variables also have an effect on the current bill rate because of the inclusion of the lagged dependent variable in the equation. The estimated effects of the policy change on the bill rate in the three quarters are 1.58, 1.59, and −2.22. In other words, the Fed is estimated to have allowed the bill rate to be higher in 1979IV and 1980I (by 1.58 and 1.59 percentage points, respectively) and lower in 1980II (by 2.22 percentage points) than it would have had it been following its old policy rule.

It is important to note in interpreting these effects that they are conditional on the lagged value of the bill rate. In 1979III, for example, the bill rate was 9.63, and given this value and the values of the other explanatory variables in equation (1) for 1979IV, the Fed is estimated to have allowed the bill rate to be 1.58 percentage points higher in 1979IV than it would have under the old rule. In 1979IV the bill rate was 11.80, and given this value and the other values for 1980I, the estimated effect on the bill rate in 1980I is 1.59 percentage points. Finally, in 1980I the bill rate was 13.46, and given this value and the other values for 1980II, the Fed is estimated to have allowed the bill rate to be 2.22 percentage points *lower* in 1980II than it would have under the old rule. (The bill rate in 1980II was 10.05.)

It should also be noted that the use of three separate dummy variables for the three quarters means that the equation is simply assumed not to pertain to these three quarters. No attempt is made here to estimate the rule that the Fed actually followed for these three quarters. The coefficient estimates of the dummy variables merely reflect the effects of whatever rule the Fed was following on deviations of the bill rate from the values implied by the old rule.

For purposes of the simulation work below, it is assumed that the Fed has gone back to its old policy rule starting in 1980III. The policy change is thus assumed to have lasted only three quarters. When more data are available, this assumption can be tested by adding further dummy variables to equation (1) and seeing if their coefficient estimates are significant. Some of the statements of the Chairman of the Federal Reserve in July 1980 are consistent with this assumption, in particular his testimony before the Senate Banking Committee on July 22, 1980.

II. The Model

The model is described elsewhere (1976, 1980b), and so it will only be briefly discussed here. The current version consists of 97 equations, 29 of which are stochastic, and has 183 unknown coefficients to estimate, including 12 first-order serial correlation coefficients. Equation (1) is part of the model. The model is non-linear in both variables and coefficients. For present purposes it has been estimated by two-stage least squares. The sample period for these estimates was 1954I–1980I except for the estimate of equation (1), where it was 1954I–1980II. The covariance matrix of the estimated coefficients, which is needed for the stochastic simulation results, was estimated using formula (4) in my article with William Parke (p. 273). This matrix, which is of dimension 183×183, is not block diagonal. Included among the 183 coefficients are the three dummy variable coefficients in equation (1).

The model has two important properties that should be kept in mind in interpreting the following results. First, interest rates have, other things being equal, a positive effect on prices. In the theoretical version of the model, which is based on the premise that firms set prices (along with other decision variables) by solving multiperiod maximization problems, the interest rate and other cost of capital variables have a positive effect on the price that the firm sets. This feature is also part of the econometric model: included among the explanatory variables in the price equation are 2 cost-of-capital variables, a bond rate and an investment tax credit variable. The second property is that prices are not very sensitive to demand changes except in periods of high economic activity. In other words, the tradeoff between output and inflation is very poor in periods of low-to-moderate economic activity. This feature, which appears to be common to many other econometric models as well, is discussed in detail in my 1978a paper.

It should also be noted that interest rates have a strong negative effect on demand and output in the model. There are a number of channels for this effect. The two long-term interest rates in the model, a bond rate and a mortgage rate, are linked to the bill rate through standard term structure equations. Both the bill rate and the mortgage rate appear directly as explanatory variables in the consumption equations, with negative coefficient estimates. Because of this, the household savings rate is, other things being equal, a positive function of interest rates. The bond rate affects prices, as mentioned above, and prices have, other things being equal, a negative effect on demand. (Prices appear as explanatory variables in the consumption equations, with negative coefficient estimates.) There is also a loan-constraint variable in the model. This variable is a function of the level of interest rates and has a negative effect on consumption in periods of tight money. Interest rates also have a negative effect on wealth in the model, through a negative effect on stock prices, and wealth has a positive effect on consumption. Demand affects output in the model, which in turn affects investment and employment; and so interest rates, by affecting demand, indirectly affect output, investment, and employment.

III. The Estimated Effects

The results for eight endogenous variables in the model are presented in Table 1. The values in the (a) rows for each variable are actual values for 1979V–1980II and predicted values thereafter. The actual values for 1980II are preliminary (they are values available as of August 1, 1980). The predicted values are from an *ex ante* forecast that I made on August 7, 1980, with the model.

The values in the (b) rows are the estimated effects of the policy change on the variables. It will be easiest to describe how these values would have been obtained had deterministic simulations been used and then to explain the modifications needed for the stochastic simulations. First, estimated residuals are available for the first three quarters (1979IV–1980II), and these residuals were added to the estimated equations and treated as exogenous. This means that a perfect tracking solution is obtained for these quarters when the actual values of the exogenous variables (including the dummy variables in equation (1)) are used. Since the predicted values beyond 1980II are based on actual values for 1980II, this also means that a simulation run from 1979IV through the end of the forecast period (1981IV) will duplicate the predicted values for 1980III and beyond, provided that the actual values of the exogenous variables are used for the first three quarters and that the exogenous-variable values used for the *ex ante* forecast are used thereafter. Call this simulation the "base" simulation.

A second simulation can then be run that is identical to the base simulation except that the values of the dummy variables in equation (1) are set equal to zero. This run is an estimate of what the economy would have been like had the monetary policy change for the three quarters not been undertaken. The difference between the values from these two runs for each endogenous variable and each quarter is an estimate of the effect of the policy change on the variable in the quarter. These differences would be the values of the (b) rows in Table 1 if deterministic simulations had been used. The modifications for the stochastic simulations will now be described.

The differences between the values from the above two runs are uncertain because they are based on estimated values of the coefficients rather than the (unknown) actual values. In a recent study (1980a), I have proposed a stochastic simulation procedure that can be used to estimate this uncertainty. The procedure in the present case consists of drawing sets of coefficient values from an estimate of the distribution of the coefficient estimates and for each set running the above two simulations. If, say, 100 draws are made, then one has 100 estimates

TABLE 1—ESTIMATED EFFECTS OF THE MONETARY POLICY CHANGE ON EIGHT VARIABLES

		1979	1980				1981				Sum over the 9
Variable		IV	I	II	III	IV	I	II	III	IV	Quarters
Bill Rate	(a)	11.80	13.46	10.05	9.27	9.16	9.13	9.12	9.16	9.24	
(percentage	(b)	1.60	2.87	0.17	0.08	0.03	−0.02	−0.06	−0.08	−0.10	
points)	(c)	0.47	0.75	0.82	0.66	0.52	0.41	0.32	0.25	0.19	
Real *GNP*	(a)	1441.7	1446.7	1414.3	1416.9	1423.1	1432.4	1443.1	1455.6	1469.2	
(billions of 1972	(b)	−1.5	−6.4	−10.3	−7.9	−7.4	−7.0	−6.5	−5.9	−5.2	−58.0
dollars)	(c)	0.7	2.6	4.8	5.2	6.0	6.7	7.1	7.3	7.2	44.7
Percentage Change	(a)	2.35	1.39	−8.66	0.72	1.76	2.65	3.04	3.61	3.78	
in Real *GNP*	(b)	−0.42	−1.40	−1.03	0.67	0.15	0.15	0.16	0.19	0.20	
(percentage points)	(c)	0.19	0.57	0.67	0.70	0.35	0.27	0.21	0.16	0.15	
Percentage Change	(a)	7.92	9.30	9.93	9.29	10.27	9.10	8.94	8.76	9.64	
in *GNP* Deflator	(b)	0.19	0.10	−0.34	0.06	0.09	−0.00	−0.02	−0.01	−0.01	
(percentage points)	(c)	0.10	0.17	0.17	0.12	0.06	0.05	0.04	0.04	0.04	
Percentage Change	(a)	7.72	0.31	0.00	8.53	8.54	8.55	8.58	8.61	8.72	
in Money Supply	(b)	−0.91	−1.35	0.03	0.09	0.10	0.12	0.13	0.14	0.15	
(percentage points)	(c)	0.33	0.46	0.51	0.44	0.32	0.23	0.16	0.11	0.09	
Private Sector											
Employment,	(a)	88253	88704	87581	86947	86843	87043	87382	87827	88359	
establishment data	(b)	−31	−159	−331	−385	−390	−390	−387	−377	−359	
(thousands of jobs)	(c)	15	70	154	209	256	302	342	372	393	
Civilian											
Unem-	(a)	5.84	6.11	7.43	8.02	8.18	8.12	7.97	7.81	7.65	
ployment Rate	(b)	0.03	0.16	0.32	0.34	0.30	0.27	0.24	0.22	0.19	
(percentage points)	(c)	0.01	0.07	0.15	0.18	0.20	0.22	0.24	0.24	0.24	
Corporate Profits,											
before tax	(a)	207.8	222.0	165.2	186.2	196.6	200.8	210.9	221.3	232.7	
(billions of current	(b)	−0.5	−5.7	−11.1	−5.7	−4.1	−3.1	−2.3	−1.3	−0.4	−34.2
dollars)	(c)	0.7	3.1	5.7	4.6	4.8	4.9	4.9	4.6	4.2	33.0

(a) rows: actual values through 1980II, predicted values thereafter.
(b) rows: estimated effects of the policy change (mean values from 150 draws).
(c) rows: standard errors of the estimated effects.

of each difference. These 100 estimates can then be used to compute the mean and standard error of each difference. For the results in Table 1, 150 draws were made, using the above-mentioned covariance matrix of the coefficient estimates for the draws. The (b) row values are the estimated means (over the 150 values for each variable and each quarter) of the differences, and the (c) row values are the estimated standard errors of the differences.

The (a) row values are subject to change in the future. Many of the "actual" values for the first three quarters will be revised, and the predicted values for the remaining quarters are not likely to be exactly right (even using my model). Fortunately, the (b) and (c) row values are not likely to be sensitive to the (a) row values, and more confidence can be placed on them than on the (a) row values. Because the model is non-linear, the multipliers in the (b) rows are a function of the (a) row values (i.e., of the initial conditions, the exogenous-variable values, and the realizations of the error terms), but for most macro-econometric models the effects of the (a) row values on the (b) row values are small relative to the size of the (b) row values.

Given the above discussion of the properties of the model, the (b) row results in Table 1 should be as expected, namely that the policy change affected output negatively but had little effect on the rate of inflation. According to the demand pressure variables in the model, the policy change was not made in a period of high economic activity. The unemployment rate in 1979III was 5.8 percent, and real *GNP* growth during the previous four quarters (1978IV through 1979III) had been only 1.8 percent. (The growth rate in 1979III was 3.2 percent at an annual rate.) The estimated effect on real *GNP* in 1980II is −10.3 billion dollars, and the cumulative effect over the nine quarters is −58.0 billion dollars. The estimated effect on the percentage change in the *GNP* deflator is −0.34 percentage points in 1980II and −0.01 percentage points by the end of the period. The rate of inflation is actually higher in the first two quarters (and in 1980III and 1980IV), which is due to the positive interest-rate effect on inflation outweighing the negative demand-pressure effect. All the (b) row values for inflation are, however, very small, and the main conclusion from them is that the policy change had very little effect on inflation in either direction.

As a consequence of the fall in output, about 400,000 jobs are lost by the end of 1980, and the unemployment rate is about 0.3 percentage points higher. The cumulative fall in corporate profits over the nine quarters is 34.2 billion dollars. The money supply grows less in the first two quarters and then slightly more for the rest of the period. Although not shown in the table, the cumulative fall in the money supply over the 9 quarters is 17.3 billion dollars (with a standard error of 13.7 billion dollars).

The standard errors in the (c) rows give one a rough idea of how much confidence to place on the (b) row values. For the first two or three quarters the standard errors are generally less than half of the estimated effects. By the end of the period they are generally greater than the estimated effects.

The standard errors in Table 1 are based on the implicit assumption that the model is correctly specified: the estimated uncertainty of the multipliers is due only to the uncertainty of the coefficient estimates. In the present case there are at least two reasons for believing that the uncertainty of the multipliers is greater than the estimates in the table. First, the model may not have captured adequately the effects of the credit controls that were imposed during part of the nine-month period. There is a loan-constraint variable in the model, and in principle this variable should have captured these effects. It may be, however, that the effects were underestimated. The decline in real *GNP* in 1980II (of 8.7 percent, preliminary estimate), for example, was considerably underestimated by the model. The model predicted (*ex post*) a fall of only 1.4 percent for the one-quarter-ahead forecast and 1.2 percent for the three-quarter-ahead forecast (i.e., the forecast beginning in 1979IV). Some of this error may have been due to a failure to capture all the effects of the controls. If so, this means that the output effects in Table 1 should be larger (i.e., more negative). The inflation effects, however, are not likely to be affected very much, given that output has little effect on inflation is this period. (The *ex post* forecasts of inflation are fairly accurate. The *GNP* deflator increased by 9.93 percent in 1980II. The one-quarter-ahead forecast was 11.38 percent, and the three-quarter-ahead forecast was 10.53 percent.)

The other reason for questioning the uncertainty estimates in Table 1 concerns the foreign sector in the model. In the current version exports and import prices are exogenous, and so foreign repercussions of the monetary policy change are not accounted for. It may be, for example, that an increase in the short-term *U.S.* interest rate results in an appreciation of the dollar (depending on how the monetary authorities of other countries respond to the increase in the *U.S.* rate). This will likely result in a fall in *U.S.* import prices and then over time to a fall in *U.S.* domestic prices. If this effect is in operation, it means that the effects on inflation of the policy change have been underestimated by the model. Some preliminary work that I have done constructing a multicountry econometric model indicate that this effect is probably small, in part because other countries' short-term interest rates respond to the *U.S.* rate.

IV. Conclusion

The main result of the simulations is easy to summarize. The change in monetary policy is estimated to have reduced real growth without having much effect on the rate of inflation. Real growth was reduced because interest rates have a negative effect on demand and output. Inflation was not affected very much because the tradeoff between output and inflation is very poor in periods of low-to-moderate economic activity. There is also an offset to the negative demand-pressure effect on inflation in this case, namely a positive interest-rate effect. The possible misspecification of the model is likely to affect the output multipliers more than the inflation multipliers. In particular, because of the credit controls, the policy change may have had a larger effect on output than is estimated in Table 1.

REFERENCES

Ray C. Fair, *A Model of Macroeconomic Activity, Volume II: The Empirical Model,* Cambridge, Mass. 1976.

______, (1978a) "Inflation and Unemployment in a Macroeconometric Model," in *After the Phillips Curve: Persistence of High Inflation and High Unemployment,* Federal Reserve Bank of Boston, Conference Series no. 19, June 1978.

______, (1978b) "The Sensitivity of Fiscal-Policy Effects to Assumptions about the Behavior of the Federal Reserve," *Econometrica,* Sept. 1978, *46,* 1165–79.

______, (1980a) "Estimating the Uncertainty of Policy Effects in Nonlinear Models," *Econometrica,* Sept. 1980, *48,* 1381–91.

______, (1980b) "The Fair Model as of August 1, 1980," mimeo., Yale Univ. 1980.

______**and W. R. Parke,** " Full Information Estimates of a Nonlinear Macroeconometric Model," *J. Econometrics,* Aug. 1980, *13,* 269–91.

MODELS FOR ANTITRUST

The Allocation of Landing Rights by Unanimity Among Competitors

By DAVID M. GRETHER, R. MARK ISAAC, AND CHARLES R. PLOTT*

During the late 1960's, air congestion often involving long delays or "stacks" was common at major airports. The right to land and take off was allocated on a first-come, first-served basis with little coordination among scheduled carriers. Since 1968, the four major airports in the United States, La Guardia, Washington National, John F. Kennedy International, and O'Hare International, have been operating under a Federal Aviation Administration (FAA) high-density ruling which limits the number of slots (takeoffs and landings per hour) at each of these airports.

Slots are allocated by scheduling committees authorized by the Civil Aeronautics Board (CAB). The scheduling committee at each airport is comprised of one representative from each airline certificated by the CAB to fly into that airport. The committees usually meet semiannually, as organized and coordinated by the Air Transport Association. Membership on the committees is relatively stable, with the same person usually being on all committees on which a carrier has representation.

The implications of the committee method of allocating airport capacity are a current policy concern. By 1985, as many as thirty-five airports may be facing serious access or capacity problems. In addition to runway, airspace, and environmental constraints, bottlenecks could be caused by loading facilities, baggage facilities, counter space, etc. Industry sources have advocated the committee process as a national solution to the associated allocation problems.

An analysis of the committee process relevant to policymakers must overcome two difficulties. First, key data about flight and route profitability will not be released by the carriers. Second, because of recent changes, the performance of the process in the past cannot simply be extrapolated to the future. Prior to deregulation, entry was effectively blocked, so the committee needed only to coordinate a few large carriers with relatively stable shares. However, with deregulation the committee must deal with entrants that seek to alter shares.

In order to deal with these problems, we studied such data as are available, and attended four scheduling committee meetings. In addition, we conducted several series of laboratory experiments.[1] The committees studied made decisions using the same procedures as do the scheduling committees. Substantial financial incentives were used to induce demand functions which had the same qualitative properties as are thought to characterize the demand functions for slots. The experimental work graphically demonstrates that the model upon which the analysis is based has empirical support. This type of evidence will probably be of little value to economists who already have considerable experience with the behavioral properties of a variety of allocation processes. The model is typical of those which are often applied, so most economists will not be surprised to see it work in a simple laboratory environment. Nevertheless, as committee processes sometimes have subtle properties, it does not hurt to check the reliability of the basic reasoning. Furthermore, some decision makers may have no experience with game-theoretic models, and rely on instincts and general theories of a completely different sort. To the extent that they may have doubts about the generality

*California Institute of Technology, University of Arizona, and California Institute of Technology, respectively. Financial support from the National Science Foundation and the Caltech Program of Enterprise and Public Policy is gratefully acknowledged.

[1] For a more complete discussion see our earlier paper.

of the economic models, additional experiments can always be conducted which incorporate the variables of their concern.

I. The Model

The model applied to evaluate the committee process is the core of a cooperative unanimity voting game without side payments. Game-theoretic models seem to provide the appropriate tools. It seems fair to say that members of the committee are aggressive defenders of their companies' interests and view the committee as a complicated bargaining process in which they apply all their negotiation skills. The value of a slot during peak hours could be worth hundreds of dollars a day. Members of the committee are generally individuals with important management positions within their companies and most have several years experience on the committee.[2] Evidence of strategic maneuvers is abundant.

The rule of unanimity captures much of the essence of the committee procedures. While the procedures used by the committee were not detailed in the order creating the committees, members were told to reach an "agreement." This has been interpreted as a basic rule of unanimity. In the past, the committee[3] has always achieved unanimity and the FAA has always approved the decision. Aside from the rule of unanimity, the committee has adopted additional procedures. Prior to each meeting the members submit their requests for slots to the committee staffs. Not surprisingly, requests for slots usually exceed the FAA quotas at least for peak periods of the day. Most of the meeting is spent in discussions among carriers and with the chair, which result in reducing the number of requests to equal the number of slots available. "Sliding," a procedure whereby a carrier moves a request for operation from one hour to another, frequently occurs. Hypothetical "exercises" are often used, with carriers constrained to the individual totals of some previous (typically the last) meeting or some other hypothetical schedule. Exercises, when complete, are usually a feasible solution which can serve as a basis for further discussion or a proposal to be voted on.

The institutional structures of the committees are designed to prevent side payments and generally induce a voting nature to the allocation process as opposed to a market nature. The committees are exempt from antitrust laws. Nevertheless, concern about potential anticompetitive effects of the committee operations led the CAB to limit the scope of the committees' activities. Each scheduling committee meeting is limited to discussions about slot allocations at a single airport for a fixed period of time. Discussions of city-pairs, scheduled fares, profitability, and other general aspects of airline competition are explicitly prohibited. Thus, for example, a committee member in the process of bargaining for an additional slot may not mention the intended destination or point of origin. These restrictions make it difficult if not impossible for the airlines to trade slots either across the high density airports or over time. Side conversations can take place but the public nature of the bargaining situation would make any "under-the-counter" sales of slots difficult. Carriers have no property rights in slots and do not have the contractual authority to make sales or trades. Carrier A may be willing to pay carrier B for slots, but if B were to reduce its slots, some other carrier (not A) may end up with them through the committee process. Thus, the institutional features suggest a game without side payments.

In all such models the core of the game is substantially influenced by the consequences of default—the option that would prevail if the committee failed to reach an agreement. No carrier would accept an allocation which it prefers less than the default option (sometimes called the "threat

[2] Clearly this has implications for the cost of this process. Meetings are held twice a year with all representatives present, and last about one week although time required has been increasing. A full four weeks were required in 1979 and most of this time was used in dealing with O'Hare and Washington National.

[3] The Washington National scheduling committee defaulted in the fall, 1980, while this paper was in press. A "temporary" allocative decision was made by the FAA and the CAB.

point" in game theory). Each member has the power to "block" group action and force the committee into default. Therefore, the final outcome must be at least as good as the default option for all members of the committee.

Should the committee fail to reach agreement, the decision would rest with the FAA. The procedure the FAA would use in the event of a default has not been decided. Four possibilities for allocating slots have been discussed: 1) a lottery; 2) an auction; 3) grandfathering slots according to historical patterns; 4) an administrative process of reviewing applications and applying some formula. No indication has been given by the FAA of its preference among these options, but carriers are not indifferent. The higher the likelihood that the FAA would grandfather slots, the less large established carriers would fear default. The higher the likelihood of a lottery or of the FAA giving slots to potential entrants, the less potential entrants would fear default.

II. Allocative Implications

An important implication of the model introduced above is that the allocations of slots within the committee processes are sensitive to the regulatory political climate. The consequences of default depend upon the decisions of the FAA which will certainly depend on the political climate at the time of default. Thus, the evaluations of the default option which are crucial from a resource allocation perspective depend in part upon political considerations.

A. *Efficiency Properties of Committee Decisions*

Allocations which result from committees using procedures such as those used by the slot committees need not be economically efficient allocations. The primary variable which guides the committee decision is the threat point (consequences of default), and given its determinants, the outcomes will be economically efficient only by accident. This general conclusion applies both at the independent committee level and at the "systems" level.

1. *Efficiency at the Single Committee Level*

The pattern has been for the new carriers to receive a few slots at the expense of carriers with a large allocation of slots. Aside from this small allocation at the time of entry, individual carriers have experienced little growth. This is understandable. Suppose the grandfather policy was adopted. The model predicts that expansion or entry could only take place if the historical time-of-day pattern was so inefficient that some carriers would prefer to give up a few slots to entrants rather than forego the gains from trade that an entrant-induced fault would cause. Thus, for practical purposes, entry and expansion would be prevented. Alternatively, if a lottery were adopted, carriers could anticipate only the expected value of the lottery. Presumably this would be the number of slots divided by number of requests where "requests" are subject to some review to avoid the obvious unbounded strategy. Without further qualifications this would mean that each carrier would expect the same number of slots. The slot committee would thus unanimously choose equal division with the largest holders forced to "give up" slots to smaller firms and entrants.

This pattern is easily seen in the experimental research. Eight (fourteen member) and ten (nine member) committee experiments were conducted with the grandfather default rule. The "historical shares" of slots across members ranged from 0 to 8 with a total of 32 and 28 units to allocate, respectively. Deviations of committee allocations from historical shares averaged only .74 slots per individual per meeting and all of this is "large" holders giving up a few slots to very small holders. By comparison, three fourteen-member committees were studied under identical parametric conditions with the exception that a lottery rule would be used upon default. All participants received either two or three slots (expected value 2.5) which is exactly the case when agents are risk neutral. Average deviation from historical share was 1.76 slots per member per meeting.[4]

[4] Using the lottery and eight grandfather experiments with identical parameters, one gets $X^2(6)=22.6$,

The current situation is probably some mixture of these two. Thus, the largest firms should be unable to expand. In fact, the largest holders should give up slots to entrants. Entrants should obtain slots until they become dubious about the default option providing them with a reasonable expectation of more.

Again the pattern is evident in the data from the controlled environment committees. Because the initial allocations need not be related to profitability, those who should expand cannot. In the controlled environment committees, there were individuals in each size class that should have grown considerably. Growth was *never* achieved for large participants and large growth was *never* achieved by smaller, nonentrant participants where efficiency demanded it. Entry was always small and unrelated to underlying profitability.

Inefficient carriers should contract in size. Certainly operations should not be transferred from more profitable applications to less. Yet the latter is what can happen within committee processes. In the experiments, for example, individuals who should have received no slots according to economic criteria always got them from the committee if the default consequences were favorable.

Economics suggests discrimination among entrants. High-cost carriers should not be granted scarce slots and enter the market when carriers with lower costs can enter or expand. Committee decisions on entry and exit do not follow this principle. There will be no exit since carriers whose operations should be replaced by other carriers have no incentive to relinquish their slots. There will also be no discrimination among potential entrants based upon their relative efficiency. All entrants have equal power to default the committee and jeopardize the slots of those who have had many. Thus, with the committee, all potential entrants can "get in." The experience of the controlled environment committees conforms to these predictions.

Given a threat point, any allocation process should exhaust "gains from trade." Generally speaking, the existing procedures are capable of dealing with that aspect of the coordination problem. The sliding operations systematically exploit the gains from trade from carriers trading operations at various times of day. The procedures are so natural that many controlled-environment committees initiated sliding operations even in the absence of their formal introduction. For the case of a grandfather default rule, efficiencies of committees that did not default always increased over the initial allocations in spite of inefficient entrants.

The sliding process does have problems. The gains from trade between two parties can be prohibited by a third member (by virtue of the unanimity rule). Thus, a member who recognizes that two other members wish to trade can use the threat of veto to gain concessions. Committee members clearly recognized this possibility in controlled-environment committees, and it appears that members of the scheduling committees also do.

2. *System Level Efficiency*

The problem of efficiency goes beyond a single airport. The value to a carrier of a slot at one airport will generally depend upon the other airports to which the carrier has access. For example, consider carriers entering a market. At a minimum this involves two airports, but because of joint costs and scale economies, entry into a "market" will frequently involve several airports. The allocation of slots within the system should be responsive to these interdependencies. The interdependencies among airports are clearly recognized by committee members.

The opportunity for some coordination across high-density airports does exist. Even though discussion of city-pairs is explicitly precluded by the initial order, references are made to other meetings. Furthermore, the meetings for different airports are often convened "back-to-back." Nevertheless the process does not seem to deal efficiently with the interdependencies. An excellent example occurred recently when TWA was willing to give up slots at O'Hare in order to increase its slots at National. United was interested in a "trade," but when other car-

which is highly significant. For this analysis, classes are defined by historical shares.

riers heard slots at O'Hare might be "released," the requests for additional slots there increased accordingly and no deal was made.

The nature of the problem is easily identified in the behavior of controlled-environment committees. For one series of experiments, payments were interdependent across two meetings. In general we found no evidence that controlled-environment committees were capable of dealing systematically with the system interdependencies.

B. *Responsiveness*

Since the committee decisions reflect primarily the consequences of default, they do not respond readily to changed economic conditions of individual carriers; indeed, they can be perverse. For example, if the profit position of a carrier increases, the optimum response in the committee can be to make *concessions* on marginal slots in order to "protect" its operations from a committee default. Thus, the firm would *contract* as it becomes relatively profitable rather than expand as it should.

More importantly, carriers do not have an incentive to replace slots when they are "unneeded" because of short term, firm specific economics. Slots released and reallocated through the committee become part of the "historical share" of another carrier and thereby affect all future decisions. Even when operations are not particularly profitable, firms have an incentive to keep them on.

C. *Susceptibility to Collusion*

Discussion of markets are strictly forbidden during committee meetings. City-pairs, prices, profits, etc. cannot be discussed. Yet, because of the committee structure, each committee member has a type of control over competitors which is uncharacteristic of markets and inconsistent with the operation of a freely competitive system. Firms can influence the *market shares among its rivals* while leaving its own constant.

As an example of these considerations consider the statement of Delta, a carrier whose position at Washington National has been very stable and thus has "given up" nothing to those who are expanding:

> I've got some numbers I'd like to read off. Postmeeting January 1978, BN had 20. Postmeeting June 1978, BN had 20. Then 22, and after the meeting last summer, BN had 24. Now with four new carriers, BN asks for 4 more, all in overage hours. I don't know whether to say congratulations or shame. *I don't intend to let BN get away with this*. I've got people who ask me about slots not being used. I explain that it's a voluntary thing, in good will. But it's harder to explain why we don't get any. I can't explain how a carrier can go from 20 to 28 (emphasis added).
> [See our paper, Appendix C]

This quotation from Delta is not atypical of concerns carriers articulate about the general slot distribution. Frequently during meetings carriers will say they will reduce requests only after "others" (often named) have done so. Sometimes they are very explicit about who they feel should get what.

D. *Long-Run Growth*

With the committee process, the value of a slot does not serve as the means and the reward for creating additional airport capacity. Instead, the slot values are capitalized in the value of the recipient carrier companies.

The committee allocation process will provide no stimulus at all for increasing airport capacity should the fiscal system fail to provide adequate funds. Or, if airport capacity is to be supplied in response to the economic demand for that capacity similar to the supply of other resources to the industry, then the committee system cannot be an adequate mechanism.

III. Recommendation

The CAB should remove the antitrust exemption of the committees. In place of the committee, we recommend the FAA establish or seek legislation which would enable the establishment of one-price sealed bid auctions with aftermarkets. The timing of the auctions and the exact definition of a slot need further study. It may also be necessary to allow provisions for "contingent bids" to deal with possibly important complementarities and nonconvexities. Revenues from the auctions should be used to relax capacity constraints. However, the exact institutional method by which the latter, important recommendation can be implemented is left for further study.

REFERENCE

D. M. Grether, R. M. Isaac, and C. R. Plott, "Alternative Methods of Allocating Airport Slots: Performance and Evaluation," prepared for Civil Aeronautics Board and Federal Aviation Administration, Polinomics Research Laboratories, Inc., Pasadena 1979.

Investment Decisions with Economies of Scale and Learniñg

By RICHARD J. GILBERT AND RICHARD G. HARRIS*

Economists in an antitrust case have at their disposal quite a large bag of tools and truisms, but for the most part these are derived from studies of static models. In many industries the policy issues concern the implications of firm behavior on market structure and performance over time. A case in point is a recent Federal Trade Commission complaint against the DuPont Corporation. The FTC alleged that DuPont had engaged in a strategy designed to monopolize the market for titanium dioxide, better known as the coloring agent in white paint. The alleged strategy was, in essence, what some would recognize as the Boston Consulting Group story: When a firm has a lead in an industry with significant learning economies, the firm should price below competitors' costs and expand to take further advantage of learning effects and, in the process, to increase market share.[1]

This paper summarizes the results of an analysis of dynamic competition with scale and learning effects. The research is a preliminary exploration. All results are obtained under rather special assumptions about the production technology, the effects of experience on costs, and the strategic interactions between firms. Two kinds of strategic behavior are considered. In the first case, each firm takes the production decisions of competitors as given (the Nash assumption). In the second case, firms consider pre-emptive capacity investments, taking into account that competitors will alter their future investment plans to achieve nonnegative profits. We call this a competitive Stackelberg (*CS*) game.

Using the maximization of net surplus as a socially optimal benchmark, neither form of competition yields an efficient outcome when new investment exhibits increasing returns to scale. In the absence of learning effects, smaller firms in a Nash competition have a greater incentive to add new capacity than do larger firms, and the equilibrium industry structure approaches equal market shares. Introducing learning effects in the Nash game causes a tendency toward increased concentration, but monopoly is not an inevitable consequence.

The *CS* game is "more competitive" than the Nash game in the sense that firms compete for the right to invest at each instant of time. With identical firms and no learning effects, the market structure in a *CS* equilibrium is indeterminate, although the sequence of industry investments is well defined. Introducing firm-specific and nonstochastic learning has a dramatic effect on the *CS* equilibrium. All new investment is undertaken by a single firm, even if the learning economies are small. Moreover, the level of output could be lower (and price higher) in a *CS* equilibrium than it would be in a Nash equilibrium. In this sense more competition can lead to a lower rate of output over time.

I. Model Structure

The model is a stylized representation of an industry which serves an expanding market and incurs large indivisible costs in installing new capacity. The technological specification is similar to that developed by Alan Manne et al. in the planning literature. This section describes the planning problem as a benchmark, but the main focus is on strategic interactions and the consequences of learning.

*University of California-Berkeley, and Queen's University, respectively. This paper is based on work reported in our earlier paper, which provides more detailed discussions of some of the propositions in this paper. We have benefited from discussions with David Newbery, Michael Spence, Robert Stoner, and Joseph Stiglitz.

[1] See Boston Consulting Group, *Perspectives on Experience*.

Assume each new plant adds one unit of capacity which can be operated at zero variable cost, and each plant is infinitely durable. Firms have access to the same technology which determines the fixed cost of building a new plant, but firms' costs may differ as a result of differential learning or experience. These knowledge effects can take many forms, each with distinct implications for market structure. This paper examines the effects of a particular learning model, where the fixed cost of a new plant depends only on the number of plants a particular firm has built. The rationale for this specification is convenience, but it serves to illustrate the implications of a particular learning process.

The solution of the planning problem yields an algorithm which determines both the optimal and the monopoly sequence of plant investments. Let $Z(k,t)$ be the flow of benefits (net surplus or profits) from k units of production capacity at date t, and assume $\partial Z(k,t)/\partial k > 0$ and $\partial Z(k,t)/\partial t > 0$ for both surplus and profits. If r is the time discount rate, the planner chooses a sequence $\{t_k\}$ of capacity installation dates to maximize

(1)
$$W = \sum_{k=0}^{\infty} \left\{ \int_{t_k}^{t_{k+1}} Z(k,t) e^{-rt} dt - C(k) e^{-rt_k} \right\}$$

where $C(k)$ is the fixed cost of the kth plant. We assume $C(k+1) \leqslant C(k)$.

The first-order conditions for problem (1), which are sufficient with the assumed restrictions on $Z(k,t)$, give the following rule for the installation dates:

$$(2) \qquad Z(k,t_k) - Z(k-1,t_k) = rC(k)$$

In the case of a social optimum, incremental surplus is set equal to the interest cost on a new plant at each construction date, while a monopolist sets incremental revenue equal to the interest cost.[2] Rapidly falling costs of new capacity may require construction of more than one plant at a particular date in a maximal program.

Remark 1: The monopoly rate of new capacity construction is slower than the social optimum.

Price is higher and output is lower than optimal under conditions of monopoly. This holds for any learning rate implicit in the cost function $C(k)$. The proof follows directly from the observation that incremental surplus exceeds incremental revenues at every date, along with the assumption $\partial Z(k,t)/\partial t > 0$.

Remark 2: The socially optimal investment path is not sustainable as a competitive equilibrium.

It is obvious that with the assumed learning function, all investment should be undertaken by a single firm. The optimum requires a monopoly industry structure, but the monopoly investment path is not optimal. Also, even if learning effects did not require a monopoly for production efficiency, the profits corresponding to the socially optimal price path can be negative and therefore the social optimum cannot be a (subsidy free) competitive equilibrium.

II. Nash Equilibrium

A theory of industry capacity investment must specify the strategic interactions among firms, and probably the most commonly used equilibrium assumption in industrial organization theory is the noncooperative game concept of Cournot-Nash. This section examines industry investment when each firm chooses a sequence of investment dates, taking as given the investment dates of others. The strategy concept is similar to the open-loop differential game, where firms choose an output path taking the output paths of competitors as given. The assumption that at the initial date agents make irrevocable commitments to actions at all future dates is not necessarily an appropriate description of competition, and the next section examines industry invest-

[2] David Starrett derives a similar characterization of optimal investment sequences.

ment in new capacity allowing for revision of firms' strategies over time.

In what follows let k represent total industry capacity. We impose the condition that capacity is fully utilized at every moment of time, which would follow from the assumption of positive marginal revenue and Cournot behavior. Firm j with capacity k^j earns revenues at date t equal to $R^j(k^j, k, t) = k^j P(k, t)$. Define the incremental revenue for firm j,

$$\Delta^j(k^j, k, t) = R^j(k^j+1, k+1, t) - R^j(k^j, k, t)$$

and assume Δ^j is a strictly increasing function of t and a strictly decreasing function of k for all k^j. The interior necessary and sufficient conditions for a Nash equilibrium are

$$(3) \quad \begin{cases} \Delta^j(k^j, k, t) = rC(k^j+1) \\ \text{if firm } j \text{ invests at date } t \text{ and} \\ \Delta^{j'}(k^{j'}, k, t) < rC(k^{j'}+1) \text{ for } j' \neq j \end{cases}$$

Our earlier paper proved this result for the case of constant capacity costs, and showed that the Nash game has strong implications for market structure.

Remark 3: If firms have access to the same technology and if there is no learning, the long-run Nash equilibrium firm sizes differ by at most one plant.

The tendency toward long-run equality of market shares in a Nash equilibrium holds for any number of firms and follows directly from the observation that the incremental revenue from a new plant is highest for the firm with the fewest plants, so the smallest firm will make the next capacity addition.

Remark 4: In a Nash equilibrium, with or without learning and with any number of firms, output is lower and price is higher than in a socially optimal allocation.

Since the incremental revenue from a new plant is less than the plant's contribution to net surplus, the Nash equilibrium investment sequence lags the optimal investment sequence. Learning economies only exaggerate this difference since unless all investment is undertaken by one firm, the firms' costs are higher than in the optimal allocation and the higher cost retards investment. If all investment is undertaken by one firm, the industry is monopolized and investment is slower than optimal (compare Remark 1).

The Nash equilibrium tendency toward equal market shares for otherwise identical firms need not hold if there is learning by doing, since learning differentiates firms according to production experience. By assumption, learning is not transferable between firms, and one might expect that this would imply a tendency toward monopoly. Although learning does imply greater concentration, strong conditions are necessary to obtain monopoly as a Nash equilibrium.

Remark 5: Monopoly is a Nash equilibrium at date T if for all investment dates $0 \leqslant t_k \leqslant T$ determined by the condition $P(k, t_k) = rC(1)$, it is true that

$$(4) \quad (k-1)[P(k-1, t_k) - P(k, t)] \leqslant rC(1) - rC(k)$$

The inequality (4) has a straightforward interpretation. The left-hand side is the loss on a monopolist's $k-1$ existing plants due to investment in a kth plant. The right-hand side is the monopolist's cost advantage relative to a firm that makes its first investment. Since a new firm does not suffer inframarginal losses from the investment, the new firm's net revenue is less than the monopolist's if the inequality is in the indicated direction. The industry is a monopoly until date T provided (4) holds for *every* plant investment in the interval $[0, T]$. The dates t_k are the possible entry dates for a new firm. It must be true that before a new firm builds the kth plant the monopolist will build the plant, and this will occur only if inequality (4) is satisfied.

The particular sequence of capacity expansions by all firms in a dynamic Nash equilibrium depends on the local properties of the cost function $C(k)$ and on the properties of the marginal revenue functions $\Delta^j(k^j, K, t)$. Although almost any sequence of investments is possible, a Nash equilibrium is always characterized by the conditions summarized in (3). The distribution of market shares can vary substantially at different moments of time; for example, one firm may be a monopoly until some date, after which several firms enter and the industry could tend toward equal market shares.[3]

III. A Pre-Emption Equilibrium

In the dynamic Nash equilibrium each firm takes the investment dates of competitors as fixed, and in this sense firms do not compete for the right to invest at each moment of time. That is, a firm could delay an investment without fear of losing a profitable opportunity because a competitor invests in its place. Some competition for the right to invest at a slot in time is inevitable, and this section describes an equilibrium with vigorous competition for investment slots.

Consider a sequence of investment dates $\{t_k\}$ and assume $k-1$ plants have been constructed. Any firm can be pre-empted from investment in slot k at t_k by any other firm if the competitor invests after t_{k-1} but before t_k. If a firm is pre-empted, it leaves the slot costlessly by not investing, but this does not preclude investing at a later date. No ties are permitted. In equilibrium the firm that invests at slot k makes nonnegative profits on the investment at t_k and no potential profits are attainable at slot k by pre-emption. This game is competitive because firms compete for investment slots and Stackelberg because the firm in slot k takes into account the possibility of pre-emption and the timing of future investments conditional on its investment at t_k.

The profit of firm j on the plant in slot k is given by

$$(5)\quad A^j(t_k) = \int_{t_k}^{\infty} e^{-r(t-t_k)} P(Q(t,t_k),t)\,dt - C$$

where $Q(t,t_k)$ is industry output at t given that the kth plant is built at t_k. In the absence of learning and with identical firms, a *CS* equilibrium can be characterized by

$$(6)\quad \int_{t_k}^{t_{k+1}} e^{-r(t-t_k)}\left[P(Q(t,t_k),t) - rC\right] dt = 0$$

Remark 6: By the definition of a *CS* equilibrium, identical firms earn zero profits, even if there are only two firms.

Existence of a *CS* equilibrium is demonstrated in our earlier paper, but the equilibrium is not unique because the sequence of plant investments in a *CS* equilibrium can be shifted over time.

Remark 7: In the absence of learning effects, output is higher (and price is lower) in a *CS* equilibrium relative to a Nash equilibrium, but the *CS* equilibrium output may be greater or less than in an optimal allocation.

The zero-profit condition implies a faster rate of capacity expansion in a *CS* relative to a Nash equilibrium. Since profits in an optimal allocation may be positive, negative or zero, the *CS* equilibrium and the optimal allocations cannot be ordered in general.

Introducing learning into the dynamic Nash game caused a tendency toward greater concentration, but monopoly was not inevitable. Learning has a much more dramatic impact on market structure in a *CS* equilibrium.

Remark 8: With learning and otherwise identical firms, all investment in a *CS* equilibrium is undertaken by a single firm. The long-run market structure with continued demand growth is a monopoly.

[3] The model of competition with learning by doing analyzed by A. Michael Spence assumes that firms take competitors' output decisions as given. Although both the Spence model and the model in Section II are dynamic Nash games, the behavioral implications differ. This follows in part from different assumptions as to the nature of the learning process.

This strong result is not dependent on the significance of the learning economies. Even a slight learning effect is sufficient to give a firm a permanent advantage with respect to investment in new capacity. To prove this result, note that in a *CS* equilibrium, if firm j invests at t_k, then $A^j(t_k) \geqslant 0$, and $A^{j'}(t_k) \leqslant 0$ for $j' \neq j$.

The term $A^j(t_k)$ is defined by equation (5), but with one complication. Since the sequence of future investment dates depends on firms' costs, and costs depend on investment experience, the output path $Q(t)$ will depend on the identity of the firm that invests at date t_k. Nonetheless, if $C(k^j) < C(k^{j'})$ (i.e., because $k^j > k^{j'}$), then for any investment sequence we have $A^j(t_k) = A^{j'}(t_k) + (C(k^{j'}) - C(k^j))$, so that $A^j(t_k) > A^{j'}(t_k)$. Therefore, firm j will always be able to pre-empt firm j'. The only question is whether it is in the interest of firm j to pre-empt firm j'. The answer is yes for the following reason. If j' invests at t_k, the additional experience lowers the cost of the next investment (and all future investments) by firm j'. The lower cost enables firm j' to make the next investment at an earlier date. This would lower the profits of firm j, and since the argument holds for every future date, it pays firm j to pre-empt firm j'.

IV. Summary

This paper has attempted to provide a structure for analyzing competition in dynamic industries with scale and learning effects. The models demonstrate the key importance of behavior in determining industry structure and performance. Aggressive competition, as illustrated by preemptive capacity investments, can lead to highly concentrated industries if a greater market share permits higher net revenues because of price coordination or firm-specific learning. This need not have negative welfare implications if the potential of entry keeps prices low and if the learning economies offset any adverse impacts from monopoly pricing. In any event, the central policy question is whether the industry allocation may be dominated by another, attainable, market outcome.

The social optimum may be eliminated as a feasible market equilibrium. Yet it is possible that rules which restrict new investment by a monopolist could increase economic surplus over the longer term. A monopoly pre-emption equilibrium is dominated by the *CS* equilibrium in the absence of learning, since the monopolist would build plants at the rate determined by potential competitors, but entry would lower prices. A policy that facilitated entry would increase net surplus provided firms continued to compete aggressively.

Perhaps more surprising, it is not too difficult to find examples where, with learning, the Nash equilibrium generates greater net surplus than the pre-emption (*CS*) equilibrium. The pre-emption equilibrium is a monopoly, and while costs are minimized, the pricing distortion can be significant. The Nash equilibrium can reduce the pricing distortion since more than one firm share the market and compete, and this can offset any reduced efficiency from lower firm-specific experience, particularly if the learning effects diminish after a few plants have been constructed.

As emphasized in the introduction, these conclusions are derived from specific assumptions and need not apply in general circumstances. Particularly suspect are the extreme conditions that learning is perfectly predictable and entirely firm specific. Learning typically occurs through a process of more or less unpredictable discoveries. Experience is retained in the minds of people who can move between firms. More study of the nature and significance of learning effects would be necessary before recommending specific policies; nonetheless, this exercise does show that learning and scale effects can have important implications for market structure and performance.

REFERENCES

R J. Gilbert and R. G. Harris, "Lumpy Investments and 'Destructive' Competition," presented at the NSF-NBER Conference on Industrial Organization and Public Policy, Berkeley, May 1980.

Alan Manne et al., *Investments for Capacity*

Expansion: *Size, Location, and Time Phasing*, Cambridge, Mass. 1967.

A. M. Spence, "The Learning Curve and Competition," work. paper, Harvard Univ., Feb. 1980.

A. D. Starrett, "Marginal Cost Pricing of Recursive Lumpy Investments," *Rev. Econ. Stud.*, June 1978, *45*, 215–27.

Boston Consulting Group, *Perspectives on Experience*, Boston 1972.

Federal Trade Commission v. E. I. DuPont de Nemours Company, FTC Docket No. 9108.

Contestability and the Design of Regulatory and Antitrust Policy

By ELIZABETH E. BAILEY*

The theory of contestable markets, as advanced in a series of recent papers, encompasses a broad variety of market forms. (See William Baumol, Bailey, and Robert Willig; Baumol and Willig; and John Panzar and Willig.) However, its most dramatic results relate to natural monopoly. The theory pertains to markets which have substantial attributes of natural monopoly, but which are characterized by free and easy entry and exit. For such markets, the cost-minimizing market structure calls for a single seller, yet the theory asserts that these sellers are without monopoly power. In the case of contestable markets, potential entry or competition *for* the market disciplines behavior almost as effectively as would actual competition *within* the market. Thus, even if operated by a single firm, a market that can be readily contested performs in a competitive fashion.

In this paper, I advance the proposition that the theory of contestable markets can be extraordinarily helpful in the design of public policy. Particular markets can readily be observed to see whether the elements of contestability are present. Even more important, the tools of public policy do, by their nature, influence the degree of contestability. Just as traditional regulatory and antitrust policies often included elements that precluded contestability, so today there is a significant opportunity to redesign public policy so as to promote contestability. The new theory can facilitate the formulation of policy which permits toleration of factors that make for natural monopoly while at the same time lessening the need for public intervention.

I. The Elements of Contestability

The key element of contestability is that a market is vulnerable to competitive forces even when it is currently occupied by an oligopoly or a monopoly. That is, if any incumbent is inefficient or charges excessive prices or exploits consumers in any other way, successful entry must be possible and profitable. Thus, in contestable markets, entry and exit must be free and easy. There can be actual competition from either within the market or from a nearby substitute service, or there can exist potential entrants who face a cost structure that is no different than that of firms already in the industry. There must also be sufficient pricing flexibility so that potential entrants can undercut current suppliers. Economies of scale and/or large fixed costs are compatible with contestability. Even markets characterized by sunk costs may be contestable if these sunk costs are readily transferable or are borne by an entity other than the firm itself. The theory of contestability does not erase all problems associated with oligopoly or monopoly supply, for there is always the possibility that substainable prices may not exist. This situation occurs, for example, when a cost structure is such that the cost of supplying a subgroup of consumers is less per unit than the cost of supplying the entire group of consumers (see Gerald Faulhaber). But in all cases where sustainable prices can exist, the theory offers strong guidance for proper design of public policy.

Intuitively speaking, the theory of contestable markets builds on the tradition of Harold Demsetz, who first pointed out that it is sunk costs not economies of scale which constitute the barrier to entry that confers monopoly power. It is primarily the risk

*Member, Civil Aeronautics Board. I wish to thank William Baumol, Nina Cornell, Vance Fort, Charles Jackson, John Panzar, Roy Pulsifer, and Robert Willig for their very helpful comments on an earlier draft of this paper. The views expressed in this paper are my own and do not necessarily reflect official policy of the Civil Aeronautics Board.

involved in expending large sums of money in order to acquire sunk-cost facilities that deters new entry when an otherwise profitable entry opportunity arises. Potential competition becomes an ever more effective force as the extent of large irretrievable entry costs declines. Similarly, incumbent firms, even those who have borne the burden of acquiring the sunk cost facility, are a problem for public policy only to the extent that they have permanent or exclusive access to that facility. Consequently, the single most important element in the design of public policy for monopoly should be the design of arrangements which render benign the exercise of power associated with operating sunk facilities.

One way to avoid the exercise of monopoly power is to have the sunk costs borne by a government or municipality, as they are in *U.S.* highway systems or airports, or by mandating that sunk costs be shared by a consortium, as is to some extent true of international broadcasting satellites, rather than to have the sunk costs incurred by the firm that is supplying the services. Virtually any method will do as long as there are contractural or other arrangements that are nondiscriminatory and permit easy transfer or lease or shared use of these cost commitments.

The theory tells us that when sunk costs are borne exclusively by a serving natural monopoly, as are railroad tracks in this country, and as are local telephone loops, then there may be a need for some form of government intervention to assure society that no excessive monopoly rents are earned from those facilities. By detaching sunk costs from the serving firm, much of the need for traditional economic regulation of the service industry disappears, even if the industry is still a natural monopoly. Instead, government intervention can often be limited to ensuring fairness of access to the sunk facility.

Fixed costs are not, according to the theory, a villian unless they also happen to be sunk. For example, although airplanes and barges might be individually costly, their mobility from market to market and their ability to be resold renders this cost unimportant as an entry barrier to a particular route and, consequently, as a source of monopoly power. Technological economies may be such that only one firm can actually serve in the market at any one time, but without exclusive rights to sunk facilities, the monopoly cannot expect to extract monopoly rents.

II. Design of Regulatory and Antitrust Policy

Unlike some policy prescriptions, the theory of contestable markets can readily be applied. The theory is clear about what types of policies enhance and what types interfere with the natural contestability of markets. In this section, I review traditional regulatory and antitrust policies which often served to preclude contestability, and describe how current policies are beginning to reverse this trend.

Consider, for example, the traditional licensing policies of the Civil Aeronautics Board (CAB) and the Interstate Commerce Commission (ICC). These policies restricted entry whether by new suppliers into the industry or by established suppliers into routes already served by others or not served by anyone. Entry was restricted both in dense markets which were structurally competitive and in thin markets where there might be expected to be only a single supplier. Authority was only conferred if it was likely to be used. There was no value placed on the benefit of having a pool of potential competitors who could respond to a potential profit opportunity by entering the market.

In contrast, the current policies of both the ICC and the CAB are to confer substantial new authority, whether actually used or not, thereby enhancing the degree of both actual and potential competition. The new policies are based on the theory that both trucking and aviation markets are, in the absence of regulatory intervention, naturally contestable. Capital is highly divisible in the trucking industry, and there is every reason to suppose that market mechanisms will work in allocating which exact commodities are carried by which particular trucking firm on which route. Even in nondense city-pair markets in aviation, where technological economies of scale with respect to aircraft size along with small traffic demand argue for a limited number of (turnaround) flights per day, potential competition should be able to act as a potent force. This is true because the major portion of airline capital costs, the aircraft, can readily be moved from one market to another. Thus, it is not surprising that John Panzar and I were able to cite evidence that, in late 1979 and early 1980 in the medium-and long-haul routes served by local service carriers, potential competition by trunk carriers was effectively policing the pricing behavior of the local carriers.

Regulation by the Federal Communications Commission (FCC) has also encouraged monopoly supply and prevented the emergence of contestable markets. There were rules that insisted upon monopoly supply of terminal equipment, even though the state of technology in that portion of the industry does not appear to favor, much less mandate the existence of a single supplier. Only in the last decade (after the Carterfone decision in 1968, the Equipment Registration Program in 1977–78, and a successful denial of AT&T's Primary Instrument Concept in 1979 by the Supreme Court), is economic control of equipment supply giving way to a policy which attempts to avoid network harm while permitting a variety of firms to supply the many terminal equipment devices that are now available for business and for household use.

Entry into the provision of transmission services was also precluded by government rule. At first, the rationale was to prevent destructive competition in an area that was considered to be a technological natural monopoly in which sunk costs played a major role. However, over time the sunk-cost aspects of cable technology have become mitigated by technical changes which led to the introduction of wireless transmission systems such as microwave services and, more recently, of satellites. With these new techniques have come new policies which have attempted to encourage both actual and potential competition. The Above 890 Decision in 1959 gave a number of moderately large firms microwave transmission privileges. The Domestic Satellites Decision encouraged firms other than AT&T to enter the domestic satellite business, and a later decision on Shared Use and Resale permitted customers of leased lines to resell their services. With the Execunet Decision in 1978, private carriers such as MCI and DATRAN were permitted to offer network services which compete in the broadest sense with the service offerings of the traditional telephone companies.

Traditional pricing policies of regulatory agencies have also interfered with the contestability of markets. One common practice has been to fix minimum or maximum rates. A typical result of this policy is that observable from stock brokerage regulation prior to 1975, where fixed rates meant that all customers paid for ancillary brokerage services, whether or not they used them. In the five years since the opening of these commission rates, a variety of price/service options have been introduced. These include discount brokers as well as arrangements whereby large buyers and sellers pay less per unit since the costs of serving them are lower.

Another practice common at the FCC and the CAB has been to set prices by formula, requiring equality of price for services over equal distance no matter whether the services involve sparsely traversed and hence costly rural routes, or routes which are heavily used and less costly. These formulas have many disadvantages. Perhaps the most obvious is the element of cross subsidy involved, with its concomitant requirement that competition be restricted on the lucrative routes. A second disadvantage of price formulas is that they tend to lag well behind technical changes. If, as in both aviation and communications, technical change tends to reduce costs more for long- than for short-haul services, the formula tends to create a bias which undervalues local services and overvalues long distance services. Yet a third disadvantage of the formulas is that they preclude price competition even where it involves innovative notions. For example, the CAB price formula of the early and mid-1970's precluded acceptance for interstate service of the high frequency/low fare proposals that were so successfully marketed by intrastate air carriers, such as Southwest Airlines in Texas and Pacific Southwest Airlines in California.

In addition, antitrust policy has been used in ways that preclude the flexibility needed for contestability. One practice has been to grant antitrust immunity for rate conferences so that competing firms set rates jointly in a government approved cartel arrangement. At the ICC, these conferences not only serve to preclude price competition, but they have resulted in the pricing of trucking services at levels which have achieved rates of return for the major carriers of 30 to 40 percent or more and substantial rates of pay for teamster members. The recently enacted Motor Carrier Act of 1980 (Pub.L. 96-296) has authorized a study commission to make a full and complete investigation upon the continued need or lack of need for continued antitrust immunity in this area.

Another practice has been to confer antitrust immunity in situations where scarce capacity is allocated. This has been done, for example, by the CAB which has granted antitrust immunity to the incumbent airlines to meet and allocate among themselves the available landing slots at four major *U.S.* airports according to a rule of unanimity. The CAB and the FAA have recently commissioned a study by the Polimonic Research Laboratories of alternative, less anticompetitive methods for slot allocation. (See David Grether, R. Mark Isaac, and Charles Plott.)

Another step in altering antitrust policy to reflect contestability theory was taken by my colleagues and myself at the CAB when we refused to use traditional market share measures to preclude mergers. In the Texas International and National acquisition case, for example, the Department of Justice recommended disapproval based in large part on market share data. They reasoned as follows: The Houston-New Orleans market shares for the twelve months ending June 30, 1978, were (in percent): National, 27, Delta 23, Texas International 24, Continental 17, and Eastern 7. The share of the two leading firms was therefore 51 percent and would be almost 75 percent after a combination of Texas International and National. This number was greater than comparable figures in mergers declared unlawful by the Supreme Court. The CAB countered by arguing that concentration ratios were not instructive in this case since with the passage of the Airline Deregulation Act of 1978 (Pub.L. 95-504), there was now relative ease of entry, even for small carriers, into such markets. In the Houston-New Orleans market in particular, there were eleven carriers with stations and functioning facilities already in place at both ends of this market. Therefore, the CAB reasoned that the markets were readily contested and did not find that a merger would be anticompetitive. Indeed, by the time the CAB order was written, a small regional carrier, Southwest Airlines, had entered the market with a low fare turnaround service and was offering approximately 25 percent of the capacity of the market. (See CAB Order 79-12-163, 164, 165.)

III. Rules for Policy Design

The previous section has highlighted a number of the principles underlying reform policies. These principles include the removal of regulatory or antitrust barriers that prevent the access of competitors or that prevent competitive pricing. They include the examination of markets to ascertain whether potential competition is workable before actual share of market is taken to be a sign of monopoly power. The theory also suggests additional rules of thumb that can be used to guide policy design.

One such rule is that there should be coordination between pricing and entry policy. Freedom of entry into a market where incumbent suppliers are constrained to price according to a regulatory formula may result in "cream skimming." Freedom to price in a market where entry is precluded by regulatory fiat may well lead to gauging of consumers. Thus, to produce results that enhance the public welfare, freedom of entry should be accompanied by freedom of pricing and the reverse.

A second rule is a smallness doctrine, under which regulatory barriers to small entrants should be removed wherever possible.

An example is the elimination by the FCC of all regulations for cable subscriber systems with fewer than 1,000 subscribers, or about 40 percent of cable systems. Another example was the decision of the CAB not to regulate route access or pricing for the commuter segment of the airline industry. Successful entry by small firms provides an excellent signal to a policymaker indicating that, under current economic conditions, there is room for enhanced competition. This competition often takes the form of the introduction of price/service innovations, such as Laker's Skytrain services from New York to London.

A third rule is that substantial pricing freedom can be granted if there is a competitive check offered by intermodal competition. It was sound policy when the ICC recently permitted railroads total pricing freedom for transportation of fresh fruits and vegetables. The ICC reasoned that a competitive check existed since truck transportation of these commodities was already deregulated. The railroads responded strongly to their new freedom and increased their market share from 24 to 40 percent in only seven months. (See Darius Gaskins, Jr. and J. M. Voytko.) Thus, by segmentation of the industry, it may become possible to permit substantial pricing freedom in areas where carriers' ability to exploit market power is curtailed.

A fourth rule for the enhancement of contestability is that entry and exit should be made as easy as possible. Expedited procedures based on written pleadings rather than oral process can enhance this process. Another idea is to shift the burden of proof so that new entrants do not have the burden of showing that entry is in the public interest, but rather incumbents must argue that it is not. Both of these policy ideas have played important roles in the reform of aviation and transportation policy.

Other rules must be devised to handle sunk-cost problems. These may include encouraging technical changes that replace technologies involving large sunk costs with technologies that offer more opportunity for mobility or shared use. They may also include a careful look by policymakers of access rules to sunk facilities. For example, access problems can arise when airport authorities attempt to meet slot or noise constraints by banning new entry while allowing incumbent carriers to expand their operations at will. They can also arise under long-term lease arrangements which allocate airport space to particular carriers, and give these carriers the power to determine when, if, to whom, and at what price to sublease space to their competitors. The problem is illustrated by Laker Airways' search for gate and terminal space at Kennedy Airport in 1977 and 1978. Because the international terminal which is owned by the Port Authority was full, Laker contacted various airlines with no success, despite the fact that at least one terminal—National's—had unused space throughout the period. Laker was unable to get help from the Port Authority. It had to sell tickets at Queens Boulevard in Long Island and take passengers and their luggage to Kennedy by bus.

In the case of railroads, one of the most difficult problems blocking comprehensive rail deregulation is associated with the costs sunk in the rail lines to major coal using facilities, such as electric power plants. Once sites have been chosen for these plants, virtually no mobility is possible. The resulting problems should be dealt with using the principles laid out in contestability theory. Policymakers should look for solutions which permit and encourage competition from other sources, such as slurry pipelines. Or policymakers could supervise a transfer of the ownership of tracks to coal mines or to the public sector, which would then seek bids for shipping the coal. The theory provides a framework for the formulation of policies capable of coping with such problems.

In communications, the sunk costs of local telephone networks are, at present, fully borne by the Bell System. At issue now is the system of prices for access to these local networks. The design of "Exchange Network Facilities for Interstate Access" (ENFIA) tariffs with the principles of contestability in mind would call for the replacement of the current negotiated

ENFIA tariffs with a system of prices that are the same for all vendors of network services so that the Bell System and others would all face the same costs of interconnection to the local network. With such equal opportunities afforded to all actual and potential competitors, with no barriers to entry, and with a policy of flexibility toward prices, the market might be expected to assure a socially efficient provision of network services.

IV. Summary

This paper has offered suggestions on how the theory of contestability can be used to organize the analysis of public policy. The theoretical concept of contestability has been shown to provide precisely the sort of guidance that has been needed for there to be an exiting confluence between economic theory and the design of regulatory and antitrust policy.

REFERENCES

E. E. Bailey and J. C. Panzar, "The Contestability of Airline Markets During the Transition to Deregulation," *Law Contemp. Probl.*, Dec. 1980.

W. J. Baumol, E. E. Bailey, and R. D. Willig, "Weak Invisible Hand Theorems on the Sustainability of Prices in a Multiproduct Monopoly," *Amer. Econ. Rev.*, June 1977, *67*, 350–65.

W. J. Baumol and R. D. Willig, "Fixed Costs, Sunk Costs, Entry Barriers, Public Goods, and Sustainability of Monopoly," in eds., William J. Baumol et al., *Contestable Markets, Industry Structure, and the Theory of Value*, forthcoming 1981.

H. Demsetz, "Why Regulate Utilities," *J. Law Econ.*, Apr. 1968, *11*, 55–65.

G. R. Faulhaber, "Cross Subsidization: Pricing in Public Enterprise," *Amer. Econ. Rev.*, Dec. 1975, *65*, 966–77.

D. W. Gaskins, Jr. and J. M. Voytko, "Managing the Transition to Deregulation," *Law Contemp. Probl.*, Dec. 1980.

D. M. Grether, R. M. Isaac, and C. R. Plott, "The Allocation of Landing Rights by Unanimity among Competitors," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 166–71.

J. C. Panzar, "Equilibrium and Welfare in Unregulated Airline Markets," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 92–95.

______ **and R. D. Willig,** "Free Entry and the Sustainability of Natural Monopoly," *Bell J. Econ.*, Spring 1977, *81*, 1–22.

Potential Competition May Reduce Welfare

By JOSEPH E. STIGLITZ*

There is a widespread belief that increasing competition will increase welfare. In general, of course, policies aimed at increasing competition will make some individuals better off and some individuals (in particular, the owners of the initial firms) less well off. The standard welfare argument is that the gainers could more than compensate the losers.

The object of this paper is to question that presumption by presenting a more general theory of market interaction in which, under quite plausible assumptions, restrictions on firm behavior which are intended to increase competition lead, in fact, to everyone being worse off.

I. The Basic Argument

Consider a market in which, at present, there is a single firm. The firm, however, is constantly subjected to competitive pressures from entry: potential rivals can engage in *R&D* activity which will result in their having a lower cost or a better product. This potential competition forces the existing firm to undertake research at a sufficient rate to deter the entry of the rival. This view of competition is more akin to the kind of competition that Schumpeter discussed than that which is conventionally presented as "pure competition."

My argument that increasing competition may lead to a Pareto inferior equilibrium is based on three critical observations: first, the amount of research in a market economy *may be* excessive. There is not a close correspondence between the returns which firms appropriate, and the social returns to *R&D*. On the one hand, it is widely recognized that there are important spillovers from any *R&D*, whether successful or not; not all increments in knowledge are patentable, and, even when a patent is feasible, other firms may be able to patent around the invention. Concern about the difficulties of appropriating returns has lead to the widespread view that there may be an underinvestment in *R&D*. At the same time, in deciding to engage in *R&D*, firms ignore the deleterious effect that a successful research program will have on the value of the capital stock of other firms. In addition, the reward provided by the patent system does not correspond directly to the *marginal* return to the particular inventor's activities, namely, the increase in the present discounted value of net benefits resulting from the fact that the invention occurred earlier than it would otherwise have occurred. The *expected* return to the inventor is the total value of the invention times the probability that he will obtain the patent (be the first to make the discovery). Under not implausible conditions, the expected private returns may thus exceed the social returns.

Secondly, the presence of potential competition alters the behavior of the existing monopolist. He will take actions to pre-empt, deter, or delay entry. These actions may lead to higher prices and lower profits, thus making both producers and consumers worse off. As discussed below, policy actions aimed at restricting some category of entry deterrent activities may lead to an increase in other entry deterring activities; again the net effect may be a reduction in profits *and* in consumers' welfare.

Finally, if the costs of production are lowered by experience (the "learning-by-doing" hypothesis), then a monopolist will

*Princeton University. This paper reports results obtained in research undertaken jointly with Richard Gilbert and Partha Dasgupta. Some of the arguments presented here are elaborated in greater detail in my referenced papers: with Partha Dasgupta, (1980a, b, 1981); with Dasgupta and Richard Gilbert, and with Gilbert. I am also indebted to Steven Salop for helpful comments and discussions. Financial support from the National Science Foundation is gratefully acknowledged.

produce beyond the point where marginal revenue equals current marginal cost of production; the firm takes into account that increasing production this period lowers production costs in subsequent periods. (See my paper with A. B. Atkinson.) Indeed, A. Michael Spence has established that, for a finite period of production and a zero interest rate, the effective marginal cost of production is exactly the marginal cost of production at the terminal period. Even if entry results in increased industry output, output per firm may be lower. If a significant part of learning is firm specific, this will increase the "effective marginal cost" of the monopolist, so that, during the period prior to entry, his output will be lower and prices higher.

More generally, markets in which there is a monopolist faced with potential entry will differ from those in which the monopolist is not faced with entry in (a) the allocation of resources to *R&D*, and consequently the date (rate) of *invention*; (b) the timing of the *introduction* of new techniques (the date of *innovation*); (c) the behavior (with respect to output, investment, and pricing) prior to the invention; and (d) the behavior after the invention. Both the pure monopoly and the monopoly faced with potential entry act in ways which are not socially optimal, but there is no presumption that one is better than the other.

The presumption that competitive markets lead to efficient resource allocations is based on the belief that there is a close congruence between private and social profits. In sectors of the economy where technological change is (or could be) important, this may not be the case, not only because of the appropriability problems which have long been recognized, but because in such markets, price will not equal marginal costs of production (otherwise the revenue required to pay for the *R&D* cannot be raised). Hence (even marginal) actions by one firm may have a significant effect on the sales (hence, profitability) of other firms. It is only under extreme and unrealistic assumptions of the Arrow-Debreu model that these externalities (sometimes referred to as pecuniary externalities) can be ignored.

II. The Natural Resource Market: An Example

Consider an intertemporal market for an exhaustible natural resource. Assume a perfect substitute for it has just been discovered which can be produced at constant costs $\bar{p}$. We contrast the situation where the substitute is controlled by the monopolist (a pure monopoly) with one in which the substitute is competitively produced. In the latter case, the substitute will be produced when the price of the natural resource rises to $\bar{p}$. The monopolist thus faces an elastic demand for the resource at the price $\bar{p}$; the presence of the competitors *raises* his marginal revenue at all dates. The presence of potential competition thus induces him, in the short run, to raise his price, although in the long run, price is lowered. For large values of the initial stock and interest rates, it can be shown that the first effect dominates the second so consumers are worse off.

There are further effects prior to the innovation. The monopolist facing the threat of entry knows that the marginal revenue, at the date of *invention*, will be higher as a result of the potential competition. This induces him to sell less prior to the invention: even before the invention occurs, the threat of potential competition raises prices.

So far, we have assumed that the date of invention is exogenous (although the date of innovation depended both on the actions prior to and after the invention). But the monopolist knows that the strongest deterrent to entry is having a large stock of natural resource; for the entrant knows that the larger the stock of the natural resource at the date of invention, the lower will prices be. Thus, the monopolist *raises* his price to reduce sales of his resource as an effective entry deterring tactic. (See my paper with Gilbert and Dasgupta.) Again, the threat of entry has resulted in consumers being worse off.

If we allow the monopolist to engage in *R&D* activity, and if there is competition for the patent for the substitute but all competitors have the same cost function, it is easy to show that the existing monopolist will *pre-empt* the competitors. But he will not charge as high a price in the period prior to the invention as when he does not engage in *R&D*. Though the monopoly is thus maintained, welfare *may* be increased. (See my papers with Dasgupta, 1980a, Gilbert and Dasgupta; and the paper by Gilbert and Newbery.) These arguments extend to more general situations with durable capital as I illustrate in the next section.

III. A Simple Model

Assume that on the date T a new firm enters a market which was previously controlled by a monopolist. (The monopolist may have a patent on a particular process for producing widgets; the new firm has just discovered an alternative way of producing the same commodity.) There is now a duopoly, and we can consider a variety of different concepts for describing the equilibrium which will emerge. For our present purposes, however, we need only note two important properties of the equilibrium. First, the profits of each of the duopolists will be a function of the *state* variables $S(T)$ describing the original firm at date T. We denote the present discounted value of the profits of the entrant (viewed as of date T) as $V^e(S(T))$ and the present discounted value of the original firm's profits by $V^m(S(T))$.

The determination of what are relevant state variables may be a fairly subtle matter. Capital goods would not be state variables, if there were perfect rental markets for them. On the other hand, long-lived durable goods which have no use other than the present one (or for which transportation costs to any other site are prohibitive) are clearly "state variables." The relevant state variables may include not only capital goods but also previously signed contracts. In the case of the learning model described briefly above, the state variable is the state of knowledge at date T. In the natural resource model, it is the stock of resource remaining at T.

Secondly, the sum of the profits are lower than what the monopolist would have enjoyed had he obtained the patent at date T. Denoting the present discounted value of these profits by $\hat{V}^m$,

$$(1)\qquad \hat{V}^m(S(T)) \geqslant V^e(S(T)) + V^m(S(T))$$

The duopoly equilibrium will not in general maximize joint profits.

Finally, denote by $\tilde{V}^m(S(T), T)$ the maximized present discounted value of the original monopolist's profits during the period $(0, T)$, given that he is constrained to have $S(T)$ at date T.

For most of the analysis, we will be concerned with situations where the date of entry of a competitor is endogenous. Assume that the entrant enters immediately after discovering an invention, which, say, lowers the cost of producing the given commodity. The present discounted value of *R&D* expenditure required to obtain the invention at date T is $R(T)$. The date of discovery, T, can be brought forward by allocating more resources to *R&D*, $R' < 0$.

I compare three market structures: 1) free entry into the *R&D* activity, but the monopolist is prohibited from engaging in *R&D* to maintain his monopoly position; 2) there is a monopolist with the same *R&D* technology facing no competition (the pure monopolist); and 3) the monopolist can engage in *R&D*, but faces competition.

A. *Pure Monopoly*

The pure monopoly problem is easiest to formulate. He simply

$$(2)\qquad \max_{\{T^m, S\}} \tilde{V}^m(S(T^m), T^m)$$

$$+ \hat{V}^m(S(T^m))e^{-rT^m} - R(T^m)$$

yielding the first-order conditions

$$(3a)\qquad \tilde{V}^m_S + \hat{V}^m_S e^{-rT^m} = 0$$

$$(3b)\qquad \tilde{V}^m_T - r\hat{V}^m e^{-rT^m} = R'$$

The firm balances off the increase in the present discounted value of profits from having, say, a larger stock of capital at date T with the cost in the obvious manner.

Similarly, it balances the gain from postponing the induced obsolence of its old capital stock with the direct gains from bringing forward the date at which the cheaper technology is introduced, and the costs of doing so.

B. *Monopoly with Threat of Entry*

The monopolist facing potential entrants but restricted from engaging in *R&D* himself) has a somewhat different problem. He maximizes

$$\tilde{V}(S(T),T)+V^m(S(T))e^{-rT} \tag{4}$$

subject to

$$V^e(S(T))e^{-rT}=R(T) \tag{5}$$

The constraint (5) reflects his belief that the *R&D* market is sufficiently competitive to drive profits to zero. Whether or when a rival enters (or engages in *R&D* in hope of discovery which will allow him to enter) depends on his beliefs about the profitability of entering. These in turn depend on the entrant's beliefs about "state" variables describing the monopolist at the date of entry. What is critical, in this view, is that actions prior to entry affect entry only to the extent that they effect potential entrants' beliefs about these state variables. High current profits or high current prices have an effect on entry only to the extent that (a) they provide *information* to possible entrants concerning the value of the relevant state variables; and (b) they affect those state variables themselves. Thus, firms with low costs may attempt to identify themselves to potential entrants by charging low prices; there may exist, as a result, a "screening equilibrium" in which all firms except the highest cost firm charge prices below the pure monopoly price, and, in more dynamic settings, may even engage in predatory pricing (charging prices below marginal costs).

Formally, the first-order conditions for the problem (4) may be written (letting μ be the Lagrange multiplier associated with the constraint (5))

$$\tilde{V}_S^m+V_S^m e^{-rT}-\mu V_S^e e^{-rT}=0 \tag{6a}$$

$$\tilde{V}_T^m-rV^m e^{-rT}+\mu(rR+R')=0 \tag{6b}$$

C. *Comparison of Monopoly with and without Entry Competition*

We now need to contrast (3) and (6). They differ in three ways:

(a) In general, $V_S^m\neq\hat{V}_S^m$. The marginal return (*after* the entry date) to capital (or some other state variable) will be different as a result of competition. In general, one might have thought it would be lower, but the natural resource model discussed briefly earlier shows that it may be higher. If competition does lower the marginal return to capital, this will lower the level of investment, and thus tend to raise prices prior to entry.

(b) In general, $T\neq T^m$. Competition will result in earlier innovation. The earlier innovation increases the rate of obsolence of capital goods installed prior to the invention, and thus lowers the level of investment, and increases the price prior to T.

My paper with Gilbert refers to these two effects as *entry accommodation*. In more general models, other entry accommodating effects may be identified. Assume the firm is uncertain about the date of entry. Since after entry, it will face more competition, it may reduce its output, though industry output may rise. It thus wants a technology with greater flexibility. The more flexible technology may have a higher average cost, and even a higher marginal cost at the point where the firm operates prior to entry. Thus prices are higher prior to entry.

(c) There is, in addition, in the presence of competition, an *entry deterrence effect* of a change in S. This leads to an increase in those state variables which discourage entry, a decrease in those which encourage entry. The effect on prices and output prior to the invention are ambiguous. For instance, assume that the durability of the capital stock is one of the variables which firms can change. Since greater durability lowers the return of the entrant, the initial firm will be induced to use machines which are more durable than he would have used in the absence of competition; since this increases his costs of production he may produce less and charge higher prices. In contrast, in the standard model (see Spence) where the only state variable is "capacity," entry deterrence leads to increased capacity and thus, presumably, to lower prices prior to entry. Similarly, in a learning-by-doing model where the state variable is the marginal cost of production at T (or equivalently, the cumulative output up to T), then lowering that will deter entry; and to lower the marginal cost of production at T, the monopolist must lower prices prior to entry. The attempt to deter entry raises welfare prior to the innovation.

Note that in many of the instances cited, the entry accommodation effects and the entry deterrence effects work in opposite directions. Thus the net effect of potential competition is in general ambiguous, and a detailed examination of the particular situation at hand is required.

D. *Pareto-Inferior Competition*

My analysis has identified two periods with differing welfare effects: Even if after the innovation T consumers are better off with competition prior to the date of innovation, consumers may be worse off in markets with potential competition, both because of the entry deterring and entry accommodation effects. In a variety of circumstances I have been able to establish that consumers are unambiguously worse off. Moreover, the monopolist is worse off, and, since *R&D* competition drives profits of entrants to zero, they are indifferent: *potential competition may be Pareto inferior.*

To see that consumers may be worse off, consider a case where T is exogenous; there is only the first entry accommodation effect. Let $\tilde{U}(S,T)$ be the present discounted value of utility up to date T, and $U^e(S)e^{-rT}$ and $U^m(S)e^{-rT}$ be the present discounted value after date T with and without competition. Let W represent the present discounted value of welfare, i.e.,

$$(7)\quad W^i = \tilde{U}(S^i,T) + U^i(S^i)e^{-rT} \quad i=e,m$$

where (W^e, S^e) and (W^m, S^m) represent the values of welfare and the state variables in the equilibrium with and without potential competition. It is immediate that $W^e \to W^m$ and $S^e \to S^m$ as $T \to \infty$. Let $\lambda = e^{-rT}$. Using (3a) and (6a) (with $\mu=0$), we observe that at $\lambda=0$

$$(8)\quad \frac{d(W^e - W^m)}{d\lambda} = U(S^e) - U(S^m) + \tilde{U}_S(S,T)[dS^e/d\lambda - dS^m/d\lambda]$$
$$= \tilde{U}_S\left[\frac{U(S^e)-U(S^m)}{\tilde{U}_S} - \frac{V_S^m - \hat{V}_S^m}{\tilde{V}_{SS}^m}\right]$$

Assume for instance that the present monopolist can produce by means of machines which are infinitely durable and which have operating costs of m_1 per unit. Each machine produces one unit and costs c_1 dollars. Let $p(Q)$ be the inverse demand function, Q_1 the output prior to entry, Q_2 the output of the entrant (under our assumptions, once he enters, his output remains unchanged), and $\bar{Q} = Q_1 + Q_2$. I assume that consumers have a separable utility function with constant marginal utility of income, $u(Q) - I$, where I represents the individual's consumption of other goods, so $p = u'(Q)$. The entrants marginal cost is m_2 and his machine costs c_2, with $m_2 + rc_2 > m_1 > m_2$. Then, at $\lambda=0$, it can be shown that

$$(9)\quad r\,d(W^e - W^m)/d\lambda = \left[u(\bar{Q}^e) - u(\bar{Q}^m) - \frac{(1-\alpha)p(\bar{Q}^e)\varepsilon(\bar{Q})\bar{Q}^e}{\varepsilon(Q_1^e)(2+\nu(Q_1))}\left(\frac{1-\alpha}{2+\nu(\bar{Q}^e)\alpha} - \alpha\right)\right]$$

where $\varepsilon = p/p'Q$, $\alpha = Q_2^e/\bar{Q}^e$, and $\nu = p''Q/p'$. The first term is positive, the second term may be negative. The second term depends on p''(i.e., u'''). There are no natural restrictions on p''. Moreover $u(\bar{Q}^e) - u(\bar{Q}^m) < p(\bar{Q}^m)(\bar{Q}^e - \bar{Q}^m)$. Since $p(\bar{Q}^e)\bar{Q}^e/p(\bar{Q}^m)$ $(\bar{Q}^e - \bar{Q}^m)$ can be very large, it is clearly possible to make the second term dominate the first, so that (at least for small λ), $W^m > W^e$: consumers are better off with pure monopoly than with potential competition. My paper with Gilbert constructed other examples where competition is Pareto inferior when the date of innovation is endogenous.

E. *Pre-Emption*

If we now allow the monopolist to engage in research, he can either not do it (thus solving problem (5)-(6)) or he can pre-empt his rivals, i.e.,

$$\text{(10)} \quad \max \tilde{V}(S,T) + e^{-rT}\hat{V}^m(S) - R(T)$$

subject to the constraint that

$$\text{(11)} \qquad V^e(S)e^{-rT} \leqslant R(T)$$

Using (1), the latter strategy dominates the former (see my paper with Dasgupta, 1980a, and Gilbert and Newbery). The first-order conditions are now

$$\text{(12a)} \quad \tilde{V}_S + e^{-rT}\hat{V}_S^m - \mu e^{-rT}V_S^e = 0$$

$$\text{(12b)} \qquad \tilde{V}_T - r\hat{V}^m(S)e^{-rT}$$
$$-R' + \mu(rR + R') = 0$$

Again, the effect on welfare of the preemption strategy is ambiguous: There is some presumption, if $\hat{V}_S^m > V_S^m$, (competition reduces the marginal return to capital) that when the monopolist is allowed to preempt his rival, he increases his investment prior to the invention. This also serves to deter entry. Thus, prices in the initial period may be lowered, but in later periods, prices may be raised.

IV. Concluding Remarks

There are two general implications of these results. First, real competitive markets are very different from the kind of idealized competitive markets modeled in the Arrow-Debreu world. The results reported here describe one attempt to provide a better basis for understanding competition in market economies. The result that increases in competition may lower welfare does not require perverse assumptions, and indeed is more general than the special models presented here; the result obtains in a variety of situations where markets are incomplete, information is imperfect and costly, and where research and development expenditures and learning are important.

Second, my results suggest that unless antitrust policy is based more directly on a more complete analysis of the functioning of competition in imperfectly competitive environments (and markets in which *R&D* are important are inherently imperfectly competitive), the pursuit of well-intentioned policies may well result in Pareto-inferior equilibria with consumers as well as producers being worse off.

Our task now is to delineate more precisely the conditions under which various policies are likely to lead to welfare improvements.

REFERENCES

A. B. Atkinson and J. E. Stiglitz, "A New View of Technological Change," *Econ. J.*, Sept. 1969, *79*, 573–78.

P. Dasgupta and J. E. Stiglitz, (1980a) "Uncertainty, Market Structure, and the Speed of *R&D*," *Bell J. Econ.*, Spring 1980, *11*, 1–28.

______ **and** ______, (1980b) "Market Structure and the Nature of Innovative Activity," *Econ. J.*, June 1980, *90*, 266–93.

______ **and** ______, "Market Structure and Resource Depletion," *J. Econ. Theory*, forthcoming.

R. Gilbert, P. Dasgupta, and J. E. Stiglitz, "Invention and Innovation under Alternative Market Structures: The Case of Natural Resources," mimeo., 1980.

______ **and D. M. G. Newbery,** "Pre-emptive Patenting and the Persistence of Monopoly, mimeo., Cambridge Univ. 1979.

______ **and J. E. Stiglitz,** "Entry, Equilibrium and Welfare," mimeo., 1979.

S. Salop, "Strategic Entry Deterrence," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 335–38.

A. M. Spence, "Entry, Capacity, Investment and Oligopolistic Pricing," *Bell J. Econ.*, Autumn 1977, *8*, 534–44.

The Revised Test of Understanding College Economics

By PHILLIP SAUNDERS, RENDIGS FELS, AND ARTHUR L. WELSH*

Like the current revision, the original development of the Test of Understanding College Economics (*TUCE*) in 1968 was a joint effort of the American Economic Association's standing committee on Economic Education and the Joint Council on Economic Education. The original *TUCE* had two main purposes: 1) to serve as a measuring instrument for controlled experiments in the teaching of introductory economics at the college level; and 2) to enable instructors to compare the performance of their students with that of students in other colleges and universitites.

Over seventy published research studies have used the original *TUCE*, and there is other evidence that the test was successful in accomplishing its objectives. But, over time, it became apparent that a revision was needed if the *TUCE* was to continue to serve effectively the purposes for which it was intended. In response to a proposal from the AEA Committee and the Joint Council, the Exxon Education Foundation agreed to underwrite the major part of the cost of revising the *TUCE*.

Our first step in the revision process was to solicit suggestions from some fifty economists who were actively involved in the teaching of introductory economics and who were familiar with the original *TUCE*. We also received suggestions from an official Advisory Committee formed to oversee the work on the *TUCE* revision.[1] It was decided to modify slightly the content specifications of the *TUCE*, and to revise considerably the cognitive specification of the *TUCE*.

I. Content Specifications

The content specifications of the revised *TUCE* are designed to reflect the subject matter emphasis of the "typical" two-semester introductory economics sequence at most *U.S.* colleges and universities. Two 30-question forms have been constructed to test five macro-economic content categories, two 30-question forms to test five micro-economic content categories, and two 30-question "hybrid" forms to test all ten content categories. Our attempts to make a clear separation of questions into macro and micro categories ran into some difficulties. While the content of most full-year courses is generally uniform, the division of topics between semesters (or quarters) is more variable. We also found that there is currently much less consensus with regard to macro content than is the case with regard to micro content.

Detailed definitions of the content categories we have used, and matrices showing the number of questions in each category, will be published in the *Manual* that accompanies the revised *TUCE*. At this point we can note that, compared to the original 1968 *TUCE* specifications, the revised macro questions give greater weight to monetary as opposed to income-expenditure considerations, and more weight to expectations and the practical problems of stabilization policy. In the micro area, the revised tests give greater weight to both market failures and

*Professor of economics, Indiana University; professor of economics, Vanderbilt University; Program Director, Joint Council on Economic Education, respectively. We are grateful to Allen Kelley for comments on an earlier, longer version of this paper.

[1] Members of this Committee are: G. L. Bach, Stanford University; Lee J. Cronbach, Stanford University; Rendigs Fels, Vanderbilt University; Allen C. Kelley, Duke University; Paul A. Samuelson, Massachusetts Institute of Technology; George J. Stigler, University of Chicago.

government failures than was the case with the original *TUCE*.

II. Cognitive Specifications

Unlike the content specifications, which sought to reflect or follow current teaching practice, the cognitive specifications of the original *TUCE* Committee sought to lead the profession in encouraging more emphasis on teaching students to *use* and *apply* economic concepts .in their introductory courses. The chairman of the original *TUCE* Committee noted:

> ...the test will emphasize the ability to apply economic principles to real problems, including issues of public policy. More specifically, the plan calls for equal numbers of three kinds of questions: (1) "recognition and understanding," (2) "simple application," and (3) "complex application."
>
> [Fels, p. 664]

The definitions of the "simple application" and "complex application" categories in the original *TUCE* proved to be somewhat difficult to interpret and use in practice. We have, therefore, revised and clarified these definitions *while still keeping the emphasis on encouraging the development of application skills*.

Our new definitions of the three cognitive categories on the revised *TUCE* (each of which has an equal weight of 10 questions on each form of the test) are as follows:

(*RU*) *Recognizes and Understands Basic Terms, Concepts, and Principles*

1.1 Selects the best definition of a given economic term, concept, or principle.

1.2 Selects the economic term, concept, or principle that best fits a given definition.

1.3 Identifies or associates terms that have closely related meanings.

1.4 Recalls or recognizes specific economic rules, for example, an individual firm's profit is maximized at that level of output at which marginal cost equals marginal revenue.

(*EA*) *Explicit Application of Basic Terms, Concepts, and Principles*

2.1 Applies economic concepts needed to define or solve a particular problem when the concepts are explicitly mentioned.

2.2 Distinguishes between correct and incorrect application of economic concepts that are specifically given.

2.3 Distinguishes between probable and improbable outcomes of specific economic actions or proposals involving no unstated assumptions or extraneous information.

2.4 Judges the adequacy with which conclusions are supported by data or analysis involving no unstated assumptions or extraneous information.

(*IA*) *Implicit Application of Basic Terms, Concepts, and Principles*

3.1 Applies economic concepts needed to define or solve a particular problem when the concepts are not explicitly mentioned.

3.2 Distinguishes between correct and incorrect application of economic concepts that are not specifically given.

3.3 Distinguishes between probable and improbable outcomes of specific economic actions or proposals involving unstated assumptions or extraneous information.

3.4 Judges the adequacy with which conclusions are supported by data or analysis involving unstated assumptions and extraneous information.

III. "Realistic" Specifications

In addition to continuing the original *TUCE*'s relatively heavy emphasis on application questions (67 percent), another dimension has been added. At least 50 percent of the application questions on each form of the revised *TUCE* are genuine or realistic applications taken or adapted from statements in newspapers, popular magazines, or other published sources.

The emphasis on realistic application questions, most of which have a fairly lengthy quotation in the stem, has caused some instructors who participated in the field testing and norming to question the extent to which we might be testing reading ability, intelligence, or sorting skill, and not economics. Hopefully, some types of economic understanding do require intelligence and the ability to read critically; but we have taken great pains to avoid "specific determiners" or questions that provide clues to "test-wise" students. It is nevertheless true that some "sorting" is required by the cognitive specifications outlined above. The Implicit Application category, in particular, tests for the ability to distinguish between relevant and irrelevant information. This may or may not be an appropriate objective in some situations, and sorting may or may not be related to economic understanding as measured by, say, recognition and understanding questions. We hope to address this issue empirically with a sample of the students in the norming groups for which we have verbal SAT scores. If student's performance on a particular question correlates more highly with their verbal SAT score than with their total *TUCE* score, or their score on a subset of *TUCE* questions, we will indeed have to take a hard look at such questions.

Before leaving the topic of cognitive specifications, we want to note that there is no *necessary* or *direct* relation between the *type of thinking* or type of knowledge being tested and the difficulty of a particular question. There are easy and hard questions in all three of the cognitive categories on each form of the revised *TUCE*. On Micro Form A the mean percent correct on both the 10 *EA* and the 10 *IA* questions is higher than the mean percent correct on the 10 *RU* questions. The mean percent correct on the 10 *IA* questions on Hybrid Form B is also higher than the mean percent correct on the 10 *RU* questions on this form of the *TUCE*, and on both Macro Form B and Hybrid Form B the mean percent correct on the 10 *IA* questions is higher than the mean percent correct on the 10 *EA* questions.

IV. Specific Examples

Two example questions from the revised *TUCE* are provided below. The data show the number of students answering the question before (pre) and after (post) they took in economics; the percentage of students choosing each alternative on each question; and the point bi-serial correlation (R) between selecting the correct alternative on each question and the total score on the *TUCE*.

Question #6 on Macro Form A is an explicit application question.

Inflation will be more difficult for the monetary authorities to contain if most people expect a rapidly rising price level, because:

A. The velocity of circulation tends to fall when the public anticipates inflation.

B. Sellers raise their prices with no regard for demand when their costs have risen.

C. Expectations of rapid inflation reduce the opportunity cost of holding money balances.

D. Expectations of rapid inflation reduce the public's willingness to hold money balances.

PRE $N=927$	*POST* $N=1163$
19%	13%
22%	11%
25%	15%
34%	61%
$R=.36$	$R=.40$

Question #17 on Micro Form A is a realistic, implicit application question.

In the early 1970's, the federal government proposed that new and stricter standards be established for sulphur dioxide emissions. Since burning coal produces large amounts of sulphur dioxide, these new standards would have especially affected coal burning firms. The president of the United Mine Workers protested the proposed standards on the grounds that they would "drive public utilities and other firms that burn large amounts of coal to nuclear reactors." This suggests that:

A. Coal was a cheap fuel partly because users could avoid some of the cost of burning it.

B. Government interference would have concealed the true economic advantages of cheap coal.

C. The sulphur dioxide standards, while well intended, were too strict to be economically practical.
D. Miners would have preferred a tax on the use of coal rather than the sulphur dioxide standards.

PRE $N=1190$	*POST* $N=1447$
29%	47%
22%	14%
36%	22%
12%	16%
$R=.34$	$R=.54$

Both of the questions shown above have good statistical properties: 1) students taking the post-test do better than students taking the pre-test; 2) all of the alternatives are plausible choices and attract some student responses; and 3) choosing the correct alternative is correlated with the total test score with a point bi-serial correlation coefficient of .34 or higher. Unfortunately not all of the questions on all forms of the revised *TUCE* possesses such good statistical properties.

V. Preliminary Analysis of Norming Data

Thirty-five different schools participated in the norming of the revised *TUCE* during the spring term of the 1979-80 school year. The schools represent a broad cross section of *U.S.* higher education institutions. Not every school participated in the norming of every form of the *TUCE*, but the post-test sample at most schools includes some students who took the same form as a pre-test, some students who took the opposite form as a pre-test, and some students who did not take a pre-test at all. We also collected information on the length of the courses (one quarter or one semester) in which the *TUCE* was administered, and on whether or not the students taking the post-test thought that their score on the *TUCE* would influence their course grade. The data shown in Tables 1 and 2 indicate that:

1) The mean post-test score is higher than the mean pre-test score on each form of the revised 30-item *TUCE* (see Table 1).

2) The post-test Kuder-Richardson Reliability Coefficients (K-R 20), which estimate consistency of measurement, are higher than the pre-test coefficients on each form of the revised 30-item *TUCE* (see Table 1).

3) As a post-test, Form B of each version of the revised *TUCE* is more difficult than Form A, and the post-test reliability coefficients are lower on Form B than on Form A for two of the three versions of the revised *TUCE* (see Table 1).

4) With one exception, the post-test difficulty of each form of the revised 30-item *TUCE* is roughly comparable to the post-test difficulty of the comparable forms of the original 33-item *TUCE* (see Table 2). The exception is revised Macro Form B which is significantly more difficult than any form of the original *TUCE*.

5) With one exception, the post-test Kuder-Richardson Reliability Coefficients are higher on each form of the revised 30-item *TUCE* than they were on the comparable forms of the original 33-item *TUCE* (see Table 2). The exception is again Macro Form B, whose K-R 20 of .76 is identical to the K-R 20 of .76 for the original *TUCE* Part I, Form B.

The original purpose of having two forms of each version of the *TUCE* was to provide one form for pre-test purposes and one form for post-test purposes. We have had problems in devising two forms of equivalent *post-test* difficulty, but the pre-test difficulty of Form A and Form B of all versions of the revised *TUCE* is similar. Instructors wishing to use one form for pre-test purposes and the other form for post-test purposes, therefore, are encouraged to use Form B before and Form A after.

The norming data indicate that the revised Macro Form A, Micro Form A, and Hybrid Form A are effective tests for the purposes that the *TUCE* is designed to serve. The data also indicate that a few questions on Macro Form B; Micro Form B, and Hybrid Form B should be revised or replaced before these instruments are used *in toto* as equivalent post-test instruments. Since our budget does not permit renorming revised forms of the *TUCE* at this time, we hope that the availability of detailed norm-

TABLE 1—PRE-TEST AND POST-TEST COMPARISONS OF STUDENT PERFORMANCE ON EACH FORM OF THE REVISED 30-ITEM *TUCE*[a]

	Pre-Test Data				Post-Test Data			
Form of Test	*N*	Mean Percent Correct	Raw Score Mean[b]	K-R 20	*N*	Mean Percent Correct	Raw Score Mean[b]	K-R 20
Macro Form A	927	36.6	10.98(3.67)	.54	1163	57.8	17.35(5.62)	.81
Macro Form B	787	35.9	10.76(3.28)	.43	1108	51.2	15.35(5.03)	.76
Micro Form A	1190	41.7	12.51(3.83)	.57	1447	55.5	16.66(4.94)	.74
Micro Form B	984	40.4	12.12(3.75)	.56	1364	55.0	16.50(4.78)	.73
Hybrid Form A	686	34.8	10.44(3.32)	.45	1084	53.7	16.12(4.81)	.73
Hybrid Form B	859	35.5	10.66(3.52)	.51	652	50.0	15.00(4.81)	.73

[a] Students and schools used for norming purposes, while similar, are not identical for various sample groups.
[b] Standard deviation shown in parentheses.

TABLE 2—POST-TEST DIFFICULTY AND RELIABILITY OF ORIGINAL 33-ITEM *TUCE* and Revised 30-Item *TUCE*[a]

	Original 33-Item *TUCE*				Revised 30-Item *TUCE*			
		N	Mean Percent Correct[b]	K-R 20		*N*	Mean Percent Correct[b]	K-R 20
Part I	FORM A	876	56.6	.76	Macro Form A	1,163	57.8	.81
Part I	FORM B	829	57.6	.76	Macro Form B	1,108	51.2	.76
Part II	FORM A	1,014	57.1	.72	Micro Form A	1,447	55.5	.74
Part II	FORM B	980	55.1	.67	Micro Form B	1,364	55.0	.73

[a] Students and schools used for norming purposes, while similar, are not identical for various sample groups.

ing data on the "good" questions on the B forms of the revised *TUCE* can be used to supplement the questions on the A forms in situations where researchers wish to vary the specifications of their measuring instruments from the fixed weights currently assigned to the new forms of the revised *TUCE*.

VI. Conclusions

The *TUCE* revision project has resulted in a set of fixed weight forms that are superior to any of the four original forms of the *TUCE*. The detailed norming data on addition questions should aid researchers in constructing variable weight measuring instruments appropriate for a variety of research purposes. In their survey article "Research on Teaching Economics," John Seigfried and Fels noted: "Future progress in economics education will require a computerized national test bank consisting of thousands of carefully edited multiple choice questions with data proving that they 'work'" (p. 929). The publication of the detailed norming data gathered in the *TUCE* revision project should represent a step forward in accomplishing this objective.

REFERENCES

R. Fels, "A New 'Test of Understanding in College Economics'," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 660–66.

J. Siegfried and R. Fels, "Research on Teaching College Economics," *J. Econ. Lit.*, Sept. 1979, *17*, 923–69.

Specification and Development of New Pre-College Tests: *BET* and *TEL*

By JOHN F. CHIZMAR AND JOHN C. SOPER*

The *Basic Economics Test* (*BET*) (see Chizmar, Ronald Halinski, and Bernard McCarney) and the *Test of Economic Literacy* (*TEL*) (Soper, 1978) are achievement tests of the basic principles of economics designed for use in grades 4 through 6 and in grades 11 and 12, respectively. The *BET* represents a substantive revision of the *Test of Elementary Economics* (*TEE*) developed in 1971, while the *TEL* is an update of the *Test of Economic Understanding* (*TEU*) published in 1963. The decision to revise the *TEU* and subsequently the *TEE* was motivated by three considerations.

First, the *Framework for Teaching Economics: Basic Concepts* (W. Lee Hansen et al.) was released in 1977. This document presented the most recent statement of the conceptual structure of the economics discipline and related that structure to decisionmaking. It was envisioned that this structure would be used by curriculum planning groups in designing K-12 economics programs. In conjunction, new evaluation instruments to be used in the planning and evaluation processes reflecting the specifications of the Master Curriculum Guide (*MCG*) *Framework* were now needed.

Second, the period of the late 1970's saw many important curriculum materials developments. Two deserve special mention. The *MCG Strategies for Teaching Economics*, available now in five grade-specific or discipline-specific volumes, have detailed specific guidelines for teaching the concepts outlined in the *MCG Framework*. In addition, a number of excellent audio-visual curriculum products have become available for use at various grade levels (see Laurence Moss). For example, *Trade Offs*, a series of fifteen 20-minute color television film programs in economics education for children 9 to 13 years of age, was released in 1978. Because of these and other curriculum developments, new testing instruments were now needed.

Third, in the decade of the 1970's many theoretical and empirical advances have been made in the models used to evaluate changes in cognitive achievement due to educational innovation (see John Siegfried and Rendigs Fels). Although these models have been devised and utilized primarily at the college level, they have begun to trickle down to evaluations conducted at the high school and grade school levels. Worthwhile evaluation must be built on a solid foundation of good data. In 1970, Wassily Leontief—in his presidential address to the Association—cautioned the discipline against elaborate model building (what Frisch called "playometrics") in the face of deficient data:

> This work can be in general characterized as an attempt to compensate for the glaring weakness of the data base available to us by the widest possible use of more and more sophisticated statistical techniques. Along side the mounting pile of elaborate theoretical models, we see a fast-growing stock of equally intricate statistical tools. These are intended to stretch to the limit the meager supply of facts.
> [pp. 2–3]

To enlarge the "supply of facts," new and hopefully improved testing instruments were clearly needed. For these reasons, the Joint Council on Economic Education (JCEE) commissioned revisions of the outdated *TEU* and *TEE*. The *TEL* and the *BET* are the results of that decision.

The first step in the revision process for both tests was the appointment of National Advisory Committees by the JCEE. For the *TEL*, the National Advisory Committee consisted of a dozen nationally recognized economists and economics educators (see

*Associate professors, Illinois State University and Northern Illinois University, respectively.

Soper, 1979, fn. 5). The National Advisory Committee for the *BET* consisted of six individuals (William E. Becker, Jr., George G. Dawson, Bonnie Meszaros, Phillip Saunders, John C. Soper, and William Walstad). These committees were appointed and organized to provide advice and consent to the Working Committees for the respective tests. The *TEL* Working Committee consisted of three economists (Michael A. MacDowell, Peter R. Senn, and John C. Soper) plus three high-school economics teachers. For the *BET*, the Working Committee was composed of two economists (John F. Chizmar and Bernard J. McCarney), an educator (Ronald S. Halinski), and four grade-school teachers.

In developing the *BET* and the *TEL*, five criteria guided the efforts of the Working Committees. First and foremost, the tests needed to reflect current thought in the discipline and current curricular thinking in economics education. Second, since it was envisioned that both tests would occupy major roles in evaluating the effect of economics education, the items that eventually would be selected for inclusion needed to be responsive to classroom instruction. Third, to minimize the effect of reading ability on the measurement of achievement in economics, the individual items needed to be written at an appropriate reading level for the intended populations. Fourth, the tests needed to meet acceptable levels of reliability for the measurement of educational achievement. And fifth, the tests needed to be power tests rather than speed tests. That is, most students should be able to complete the tests in a 50-minute class period.

It became apparent to the Working Committees throughout the stages of development of both tests that these criteria were not totally compatible. For example, the use of economics jargon and even the attempt to "write around" such terms was at odds with the reading-level criterion; the inclusion of items responsive to instruction, in many cases, meant the selection of items which did not necessarily maximize reliability; the attempt to construct power tests placed constraints on the number of economic concepts which could be tested individually; and attempts to insure "content validity"—the extent to which the tests covered what the National Advisory Committees determined *ought* to be tested—sometimes conflicted with basic curricular considerations and therefore with responsiveness to instruction (i.e., what *is* taught is not always what disciplinary experts think ought to be taught). To continuously check our judgment on issues such as these, our mode of operation was to seek input from members of the National Advisory Committees as often as possible. A description of the developmental process follows.

A test matrix, one dimension being economic content and the other being level of cognitive functioning, was developed for each test and used as a guide in determining what was to be included in the instruments. The purpose of these specification matrices was to ensure the content validity of the actual tests. The specifications also provide teachers who might have different objectives the information necessary to interpret results of the tests correctly.

For both the *BET* and the *TEL*, the content categories are based upon the *MCG Framework* (see Hansen et al., pp. 5, 8, 9). It was decided to eschew the classification scheme for cognitive functioning used in the Test of Understanding in College Economics (*TUCE*), i.e., recognition-understanding, simple application, and complex application. The Working Committees for both tests concluded independently that the *TUCE* schema was unfamiliar to most individuals (mostly teachers) who would be likely to use the tests and was not operational (in that it was difficult, if not impossible, to achieve committee consensus as to the proper classification of any given item). Instead the *BET* employs three cognitive categories (knowledge, understanding, and application) based on a modified taxonomy (based on Benjamin Bloom). The *TEL* uses a five-level taxonomic categorization (knowledge, comprehension, application, analysis, and evaluation) again based on Bloom. Because Bloom's taxonomy—which is widely known among potential test users—is developmental over time, the Working Committee for the *BET* decided that knowledge, under-

TABLE 1—ORIGIN OF *TEL* ITEMS
(Forms A and B Combined)

			Average *P*-Levels Original Tests		Average *P*-Levels *TEL*	
Source	Number of Items	Percent of *TEL*	Without Economics	With Economics	Without Economics	With Economics
TEU:						
exact copies	7	7.6	57.4	71.9	47.5	58.9
minor changes	13	14.1	52.5	65.8	43.7	55.3
major changes	25	27.2	44.3	56.6	38.9	49.4
Total *TEU*	45	48.9	48.7	61.6	41.6	52.6
TUCE	9	9.8	–	73.9	39.5	49.5
Other Tests	7	7.6	41.3	47.1	60.2	66.9
New Items	31	33.7	–	–	41.7	50.0

standing, and application of the basic economic concepts are what can reasonably be expected of 4th-6th graders. Thus, the *BET* is weighted in favor of these cognitive outcomes. The *TEL* Working Committee decided on an expanded taxonomy including analysis and evaluation items which are appropriate for high-school juniors and seniors.

From this point the developmental processes for the two tests diverged somewhat, largely because the *TEL* Working Committee concluded that a large number of questions from the original *TEU* (Bach, Jones, and Meyer) could be used in the new instruments with little or no change. Analysis of the *TEL* specification matrix also pointed up a number of areas in which questions would have to be either drawn from other testing instruments or constructed anew. Table 1 displays the genesis of the 92 items contained in the final two forms of the *TEL*. A total of 45 *TEU*-based items were used, 7 of which are exact copies of *TEU* items. Thirteen old *TEU* items required minor wording changes, while 25 items involved major changes—often a complete rewrite. In addition, 9 items were drawn from the *TUCE* (5 of which being exact copies), 7 items came from other tests (the *Junior High School Test of Economics* and the *Test of Understanding in Personal Economics*), while 31 items had to be written "from scratch."

Table 1 also indicates the percentage weights on the *TEL* for each class of question by origin and the average *P*-levels (percentage of the relevant norming populations answering the items correctly) from the norming of the original instruments and the *TEL*. A comparison of the *TEU* and *TEL* *P*-levels reveals substantial differences in the performance of student groups on the same items at different points in time. Such a comparison suggests either that the *TEU* norming sample was drawn from relatively high-ability students or that the average high-school student knew significantly less economics (or was of lower general ability) in 1977 than was the case in the mid-1960's when the *TEU* was normed.

For the *BET*, it was decided that all new questions would have to be developed to fill out the *BET* specifications. The Working Committee generated a large number of items which were critiqued for content, readibility, ambiguity, and general relative worth. The items were placed into the text matrix according to their concept and behavior classification. From these items, prenorming versions of the test were constructed for initial tryout. These tests were administered to children in all the grades of the relevant population, both with and without economic instruction, and organized such that each form was administered in every class. Trial versions of the *TEL* (four in all) were similarly administered to sample populations of high-school students to generate debugging data.

The prenorming versions of the *BET* and the *TEL* were analyzed for item difficulty, reliability, and responsiveness to instruction. The analysis of the data also helped to detect flaws of exposition that might act to

TABLE 2—DESCRIPTIVE STATISTICS FOR THE BET AND TEL

		$\bar{X}$	σ	N	SEM	α
BET	Form A	18.69	6.49	7,031	2.69	0.829
	Form B	18.56	5.84	6,777	2.71	0.784
TEL	Form A	21.59	8.52	4,192	3.02	0.875
	Form B	22.89	8.43	4,468	3.01	0.872

confuse or mislead students. But pre-testing does more than illuminate flaws in item exposition. It also reveals which questions have the power to discriminate. On the basis of data derived from the trial administrations, two forms of each test were developed for norming purposes and distributed to the National Advisory Committees for final comments.

Based on socioeconomic data previously collected, a large number of tests were mailed out to obtain data on student performance both with and without prior economics instruction. We cannot guarantee that the norming populations for the two tests were exactly representative of the student population enrolled in schools throughout the nation, but the procedures followed suggest that the norming groups were reasonably representative of school districts demonstrating interest in economic education.

Once again, statistical data from the norming populations were obtained on item difficulty, discrimination power, reliability, and responsiveness to instruction. This led to the deletion of two items on each form of the *BET*. The *BET* forms were then rescored and reanalyzed. For the *TEL*, it was decided to retain all 92 items used in the norming version. The final forms of the *BET* and the *TEL* reflect the judgment of the respective Working Committees in balancing the demands of the various criteria we attempted to meet, as well as the sometimes conflicting feedback received from different sources.

The purpose of administering the norming versions of the test instruments was to obtain data to assist users in making meaningful interpretations of test scores. The mean, standard deviation, sample size, standard error of measurement, and Cronbach *Alpha* reliability for each form of the *BET* and *TEL* are reported in Table 2. Detailed analysis of both tests can be found in the test manuals (see Chizmar and Halinski; Soper, 1979).

What can the new tests accomplish? Before answering this question, a point of conflict between the two authors of this paper requires explanation. Chizmar feels that teachers should not use the *BET* as a pedagogic aid in teaching the very concepts the test was ostensibly designed to examine. (Because of the inclusion of the "item rationale" section in the manual, this becomes a very real possibility.) To do so would destroy the specifications of the test matrix upon which the test was built. Once students have been exposed to the questions in a teaching session, every item becomes a "knowledge" item, no matter how it was originally classified. Thus, the validity of the test would be destroyed.

Soper's position on this issue is that the *TEL* may be legitimately used as a powerful teaching tool. While acknowledging the danger of "teaching to the test," he argues that teacher-led discussion of the individual test items (particularly following post-testing) can be used to clarify and expand student knowledge of the concepts the students ought to understand. If alternate forms of the test are used in pre- and post-test settings, no violence is done to the test specifications as a result of classroom discussion (and teaching) of the specific test items. We leave the resolution of this conflict to the individual user of either test.

Beyond such differences, the authors feel that the new tests provide the nation's schools with an updated set of tests to evaluate curriculum development. Such evaluations can occur in a number of settings. There can be a formal school district curriculum for instruction in economics covering more than one grade. There can be no formal curriculum for instruction in economics but where a school district wishes to monitor student growth in economics taking place through informal learning. Finally, the tests can be used where economics instruction is primarily an individual teacher project carried out somewhat informally. Whatever the setting, used properly these instruments will provide information concerning student growth and achievement, and the effectiveness of educational materials and teaching strategies.

Finally, the new tests are and will continue to be research instruments. The *BET* and the *TEL* were designed to measure the degree to which economics instruction aids in understanding and using the concepts outlined in the *MCG Framework*. Certainly, many exciting curricular developments have occurred recently. These innovations cry out for evaluation. Hopefully, these new precollege tests will assuage the "meager supply of facts" now available.

REFERENCES

George Leland Bach, Walter R. Jones, and Suzanne R. Meyer, *Interpretive Manual and Discussion Guide: Test of Economic Understanding*, Chicago 1964.

Benjamin S. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain*, New York 1956.

John F. Chizmar and Ronald S. Halinski, *Basic Economics Test: Examiner's Manual*, New York 1981.

______, ______, and Bernard J. McCarney, *Basic Economics Test: Forms A and B*, New York 1980.

W. Lee Hansen et al., *Master Curriculum Guide in Economics for the Nation's Schools; Part I: A Framework for Teaching Economics: Basic Concepts*, New York 1977.

W. Leontief, "Theoretical Assumptions and Nonobserved Facts, "*Amer. Econ. Rev.*, Mar. 1971, *61*, 1–7.

L. S. Moss, "Films and the Transmission of Economic Knowledge," *J. Econ. Lit.*, Sept. 1979, *27*, 1005–119.

J. J. Siegfried and R. Fels, "Research on Teaching College Economics: A Survey," *J. Econ. Lit.*, Sept. 1979, *27*, 923–69.

John C. Soper, *Test of Economic Literacy: Forms A and B*, New York 1978.

______, *Test of Economic Literacy: Discussion Guide and Rationale*, New York 1979.

Master Curriculum Guide in Economics for the Nation's Schools; Part II: Strategies for Teaching Economics, 5 vols., New York 1978–80.

Trade Offs, 15 vols., Bloomington, Indiana 1978.

Social Welfare Dominance

By ROBERT D. WILLIG*

A social state is said to *social welfare dominate* another if it would be preferred by all social decision makers with preferences that satisfy a specified set of axioms. The aims in taking this approach to normative analysis are to specify axioms that represent widely appealing social values, and to derive useful characterizations of the corresponding social welfare dominance partial ordering of social states. The axioms described below obviate both neglect of income distributive concerns and arbitrary specification of a particular Bergson-Samuelson social welfare function. Yet, they yield a social welfare dominance relation that is more conclusive than Pareto dominance, and that reflects both equity and efficiency concerns.

In Section I, I present the axioms and a characterization in terms of real incomes of whether one social state social welfare dominates another. In Section II, I detail the components of the requisite concept of real income. In Section III, I sketch some applications of these normative methods, and finally, in Section IV, I begin the discussion of the ethical contents of social welfare dominance.

I. Social Welfare Dominance

For this brief exposition, I work in a simple framework. The analysis applies to a fixed population of n neoclassical households. The welfare of the ith household is represented by its ordinal indirect utility function, $l^i(s_i, p, m_i)$, where s_i is a vector of selected household (or environmental) endowments (or characteristics), p is the vector of prices, and m_i is the household's income. A social state ρ is completely summarized by the endowments of each household, the price vector, and the income of each household: $\rho=(s, p, m)$.

The axioms that I utilize on the preference orderings of social decision makers (*sdm*) over social states are as follows:[1]

A1: (Pareto Principle) The *sdm* is indifferent to a change that leaves all households indifferent and does not prefer a change that no household prefers. For at least one household, a change that is preferred by it and that leaves all others indifferent is preferred by the *sdm*.

A2: (Anonymity) If all households faced the prices p^0 and had the standardized levels of selected endowments s^0, then the *sdm* would be indifferent to a reversal of nominal incomes between any two households.

A3: (Regressive Transfer Aversion)[2] If all households faced the prices p^0 and had the standardized levels of selected endowments s^0, then the *sdm* would not prefer any lump sum transfer of nominal income from a nominally poorer to a richer household.

Axiom A1 is a standard, relatively weak, form of the Pareto principle that incorporates "respect for individual (household) preferences." Axioms A2 and A3 stipulate features of the preferences that the *sdm* would exhibit with respect to the distribution of nominal income under particular (and possibly counterfactual) circumstances. These circumstances are that particular *base prices* p^0 prevail, and that all households

*Princeton University. The research described here was supported, in part, by the National Science Foundation. Portions of this paper draw on joint work with James McCabe. Direct precursors are Anthony Atkinson; Amartya Sen; Partha Dasgupta, Sen, and David Starrett; Michael Rothschild and Joseph Stiglitz, and my paper with McCabe. See my article with Elizabeth Bailey for more detailed expositions of some of the material here, for some pilot applications, and for more complete references. I am grateful to E. Bailey, J. Chipman, A. Klevorick, G. Loury, and R. Schmalensee, among others, for helpful feedback.

[1] These axioms extend those in my paper with McCabe, which are themselves extensions of those in Rothschild and Stiglitz.

[2] This term was suggested by Alvin Klevorick.

have the same base levels s^0 of the household (or environmental) endowments (or characteristics) represented in the vectors s_i. The axioms do not require standardization of whatever other household characteristics are subsumed in the utility functions themselves.

I argue below that there is substantial ethical content in the choices of the endowments or characteristics to be standardized, of s^0, and of p^0. Axiom A2 stipulates that the *sdm* would view households as equally deserving, except inasmuch as their nominal incomes were different, if they all faced the base prices and had the base levels of the standardized endowments. Axiom A3 stipulates that, under these same circumstances, the *sdm* would not prefer any change that entails nothing more than a lump sum regressive transfer. It is this axiom that injects income distributive concerns into social welfare dominance.

Definition: The state ρ' *social welfare dominates* the state ρ if ρ' is preferred to ρ by all *sdm*s whose preferences satisfy axioms A1, A2, and A3.

The statement of the characterization theorem requires this definition.

Definition:[3] The income compensation function $\mu^i(s^0, p^0 | s_i, p, m_i)$ is the nominal income that household i would require, with endowment levels s^0 and prices p^0, to be indifferent to nominal income m_i, with endowment levels s_i and priees p. This is the household's real income, base (s^0, p^0), when it actually has (s_i, p, m_i).

THEOREM:[4] *The state* $\rho' = (s', p', m')$ *social welfare dominates the state* $\rho = (s, p, m)$ *if and only if*

$$(1) \qquad \sum_{i=1}^{k} Z'_i > \sum_{i=1}^{k} Z_i, \quad k = 1, \ldots, n$$

where $Z'_i \equiv \mu^i(s^0, p^0 | s'_i, p', m'_i)$, $Z_i \equiv \mu^i(s^0, p^0 | s_i, p, m_i)$, *and where the indexes* i *are assigned to possibly different households under* ρ' *and* ρ *so that* $Z'_1 \leqslant Z'_2 \leqslant \ldots \leqslant Z'_n$ *and* $Z_1 \leqslant Z_2 \leqslant \ldots \leqslant Z_n$.

The theorem says that a change is social welfare dominating if and only if it raises the real income of the poorest household, raises the total real income of the poorest two households, raises the total real income of the poorest k households $(k = 1, 2, \ldots, n)$, and raises the total real income of all households. Thus, both the Rawlsian and Hicksian dictates are reflected. *A dominating change can lower the real income of a group of households, but it must raise the total real income of all poorer households by more.*

II. The Assessment of Real Incomes

As defined above, real income is measured relative to base prices p^0 and to the base levels s^0 of the standardized endowments. It is useful to decompose real income (Z_i) into three parts: nominal income $= m_i$; an equivalent variation for prices $EV_i \equiv \mu^i(s_i, p^0 | s_i, p, m_i) - m_i$; and a compensating differential for endowments, $CD_i \equiv \mu^i(s_i, p^0 | s_i, p, m_i) - \mu^i(s^0, p^0 | s_i, p, m_i)$. By these definitions, $Z_i = m_i + EV_i - CD_i$ and $\Delta Z_i = \Delta m_i + \Delta EV_i - \Delta CD_i$.

EV_i is the standard equivalent variation for price changes, defined here relative to the household's actual endowments, s_i. As such, both EV_i and ΔEV_i can be calculated from estimated demand systems, and approximated by consumer's surplus (as in my earlier article). The lower are prices p, the higher are both EV_i and real income Z_i.

CD_i is the change in the household's measured real income that is caused by changing the base from the household's actual to the standardized levels of endowments. Also, $CD_i = \mu^i(s_i, p^0 | s^0, p^0, Z_i) - Z_i$, the compensation the household would require for accepting its endowment levels in place of the standardized ones, at base prices and income Z_i. Thus, the less desirable are a household's endowments, the larger is its compensating differential and the smaller is its real income, Z_i. For marketable endow-

[3] This is a simple extension of the definition created by Leonid Hurwicz and Hirofumi Uzawa.

[4] A similar result appears in my paper with McCabe. The arguments given in Rothschild and Stiglitz can be modified to prove this theorem. A direct proof is given in my article with Bailey.

ments, CD_i is simply the outlay flow required at base prices to exchange the household's levels for the standardized ones. However, the quantitative assessment of CD_i would become less straightforward if, for example, health, abilities, age, or household composition were included in the list of selected endowments and charàcteristics.

Nevertheless, the conditions for social welfare dominance would be particularly tractable if: (i) the rankings of households by real income were the same in the two states; and (ii) $\mu^i(s_i, p^0|s^0, p^0, y) - y$ were invariant in y. Then $CD_i' - CD_i = 0$, $Z_i' - Z_i = \Delta m_i + \Delta EV_i$, and the investigation of (1) could proceed with standard tools. Although there would be no ΔCD component in real income *changes*, the compensating differentials could play key roles in determining the real income *ranking* of households. This ranking could be defined without recourse to *formal* evaluations of all of the selected household endowments and characteristics.

III. Methods of Application[5]

The characterization of social welfare dominance given by the theorem is unwieldly inasmuch as (1) directs the normative analyst to check inequality relationships equal in number to the population of households. Fortunately, several devices may render the requisite analysis more tractable. First, in some circumstances, it can be appropriate to work with a relatively small number of groups of households with homogeneous levels of real income.

Second, in general, (1) is equivalent to $Z_1' - Z_1 > 0$ and $\Sigma_{i=1}^{k}(Z_i' - Z_i) > 0$ for only $k = n$ and those k for which $Z_k' - Z_k < 0$ and $Z_{k+1}' - Z_{k+1} \geqslant 0$. That is, with $Z_i' - Z_i$ plotted against i, or Z_i, the inequalities in (1) need only be checked for $k = 1$ and $k = n$, and where the curve crosses zero from below. This can be an enormous simplification, especially where normative analysis can be based on a cross-sectionally estimated complete demand system.

As the simplest example, suppose that all households have identical preferences, albeit various incomes, and that there are no selected standardized endowments in axioms A2 and A3. To compare two social states that differ only in prices, the tests given by (1) are whether $\int_a^y \Delta Z(m)\,dF(m) > 0$ for $a < y \leqslant b$, where $\Delta Z(m) \equiv \mu(p^0|p', m) - \mu(p^0|p, m)$, a and b are the minimum and maximum household incomes, $F(m)$ is the cumulative income distribution, and μ is the common income compensation function. These tests are equivalent to checking the sign of the integral for y near a, $y = b$, and for intermediate values of y at which the $\Delta Z(y)$ curve crosses 0 from below. It is easy to see that if the estimated complete demand system has linear Engel curves (i.e., the linear expenditure system or any other Gorman normal form), then $\Delta Z(y)$ is linear in y and can thus cross zero only once. Also, for the translog indirect utility function, $\Delta Z(y)$ can cross zero only once. In any such case, tests at intermediate values of y are unnecessary. Thus, there, price changes are social welfare dominating if and only if they raise total real income and the real income of the poorest household.

The third device for simplifying analysis of (1) is the following result: It suffices for (1) that it hold for $k = n$ and that B_i / C_i be decreasing in i, where $\Delta Z_i = B_i - C_i$, $B_i > 0$, and $C_i > 0$. *A change that is efficient in that it raises total real income is social welfare dominating if the associated class-specific benefit-cost ratios decline with real income.* This result can be useful where qualitative information indicates the relationship between real income levels and benefit-cost ratios. For example, a change comprised of a price increase and a price decrease has associated benefit-cost ratios that are related to the ratios of demand for the two goods. It follows that the pair of price changes is social welfare dominating if it is efficient and if the cross-sectional (real) income elasticity of demand for the good whose price is increased exceeds that of the good whose price is decreased. Of course, it

[5] This section is largely drawn from my paper with McCabe.

is comforting to thus discover that an increased sales tax on diamonds, together with a decreased sales tax on food, would be social welfare dominating if it were efficient.

IV. Ethical Content

The ethical content of social welfare dominance can be explored through examination of its underlying axioms, as well as through its implications. Axioms A2 and A3 require that households would be viewed as anonymous, except for their incomes, if they faced base prices and they shared the standardized levels of selected endowments. This last, complicating, stipulation may well be necessary for the axioms to reflect widely held social values.

Otherwise, for example, A2 would require indifference to an interchange of nominal incomes between a healthy household and a household with a larger income and severe physical disabilities. Similarly, A3 would proscribe preference for a transfer from a healthy slightly poorer household to a severely disabled slightly richer one. However, with health status one of the selected endowments for standardization, A2 would only require indifference to interchanges of nominal incomes under the circumstances that all households exhibited the same base level of health status. Yet, the axiom would still require indifference to interchanges of nominal incomes, provided that they were accompanied by appropriate compensating differentials for the households' actual levels of health status. And A3 would permit preference for the aforementioned transfer if the relevant compensating differentials reversed the real income rankings of the households.

Correcting nominal incomes for variations in such endowments as health status and material wealth, in such a fashion, would certainly seem to be consistent with prevalent social values. Obversely, prevalent *U.S.* public ethics seem to require that public treatment of households be unaffected by their levels of such endowments and characteristics as personal relationship to the *sdm*, physical beauty, and religion. Thus, social values would exclude these endowments from the selected set to be standardized, and would include them under the umbrella of anonymity.

It would seem that *U.S.* opinion is divided on whether anonymity should reach such other household endowments and characteristics as race, household composition and size, drug addiction, farm land ownership, marital status, and age. Controversies abound over government policies and programs that accord differentiated treatments to households on the basis of such descriptors. I feel that much insight could be gained by recasting some of the ethical issues that underlie these controversies into a well-defined form of question: Should a given household descriptor be utilized as a basis for adjusting real incomes with compensating differentials, before equity precepts are applied in an anonymous manner?

However, at least one caution should be noted here. Some characteristics that are candidates for special treatment may be partially under the control of the household (for example, family size, marital status, unemployment and drug addiction). Then, preferential programs and policies may stimulate the very problems they aim to alleviate. As such, some controversies superficially cast as debates over ethics may, instead, fundamentally concern tradeoffs between efficiency and equity. While issues of this kind can be studied within the social welfare dominance framework, they do not concern the selection of endowments for standardization in the axioms.

The choice of the price vector to be utilized as a base in the axioms is also a matter of substance for social welfare dominance. The partial ordering characterized in the theorem can depend on p^0, and, in constructed examples, alterations in p^0 can even cause dominance relations to reverse.[6] Thus,

[6] However, alterations in p^0 that leave unchanged the identity of the poorest household cannot reverse a dominance relationship. Further, the consumer's surplus representation of ΔZ_i shows that moderate alterations in p^0, that leave the real income ranking of households unchanged, are unlikely to perturb the

like other normative methods, social welfare dominance is not immune to index number problems. However, in this context, at least some of the sensitivity to base prices that the normative comparisons exhibit may be ascribed to ethical implications of the choice of base prices, rather than to meaningless ambiguity in the methodology.

For example, consider the choice of the base level of the price of heroin. The smaller it is, the lower will be the relevant levels of the real incomes, Z, of heroin users relative to those of households which abstain. However, if the base price of heroin were equal to the market price, then the real incomes of all households would incorporate no adjustment for the level of demand for the drug. In this case, taste for heroin would be an anonymous characteristic, and axioms A2 and A3 would preclude any special consideration for heroin purchasing households.

On the other hand, stipulation of a small base price of heroin in A2 would require that the *sdm* treat households in an anonymous fashion only if the price of heroin were suitably low. Given a higher market price for the drug, the relevant real incomes of heroin using households would be adjusted downward by their equivalent variations (EV_i), before the anonymity and regressive transfer aversion values were applied to them. Thus, the choice of a small base price for heroin would reflect the ethical stance that heroin users merit special consideration and should be viewed as poorer than their nominal incomes would indicate. Obversely, choosing a base price for heroin near market levels would reflect the ethical stance that demands for the drug should be viewed no differently, for equity purposes, than the demands for any other commodity.

The example indicates why I feel that the sensitivity of the social welfare dominance relationships to the choice of base prices is neither an index number *problem* nor a methodological ambiguity. Rather, I think it reflects the fact that the choice of base prices, like the selection of the standardized household endowments and characteristics, embodies substantial ethical content.

The axioms presented here provide a framework for formal expression of a rich class of social values and ethical positions that are not otherwise readily represented. Further, the framework enables formal analyses of the implications of such values and ethics for normative assessments of social states and policies. Moreover, the theorem provides a new practical method for welfare analysis that incorporates both concern for economic efficiency and well-defined concern for distributive equity.

dominance partial ordering because they little effect each ΔZ_i. Moreover, the dominance partial ordering over price vectors is invariant to base prices in the context of a cross-sectional complete demand system with the Gorman normal form.

REFERENCES

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

P. Dasgupta, A. K. Sen, and D. Starrett, "Notes on the Measurement of Inequality," *J. Econ. Theory*, Apr. 1973, *6*, 180–87.

L. Hurwicz and H. Uzawa, "On the Integrability of Demand Functions," in John S. Chipman et. al., eds., *Preferences, Utility, and Demand*, New York 1971, 114–48.

J. McCabe and R. D. Willig, "Consumer's Surplus and the Effect of Relative Price Changes on the Distribution of Individual Real Income," unpublished manuscript.

M. Rothschild and J. Stiglitz, "Some Further Results on the Measurement of Inequality," *J. Econ. Theory*, Apr. 1973, *6*, 188–204.

Amartya Sen, *On Economic Inequality*, Bath 1973.

R. Willig, "Consumer's Surplus Without Apology," *Amer. Econ. Rev.*, Sept. 1976, *66*, 589–97.

______ **and E. Bailey,** "Income Distributional Concerns in Regulatory Policy-Making," in Gary Fromm, ed., *Economics of Public Regulation*, forthcoming.

Arbitration and Conflict Resolution in Labor-Management Bargaining

By VINCENT P. CRAWFORD*

Compulsory arbitration is frequently employed to resolve labor-management bargaining disputes when the union is legally prohibited (as are, for example, many public employees' unions) from striking. In this form of arbitration, an arbitrator is empowered to impose a settlement on the bargaining parties if their negotiations break down. Various compulsory-arbitration schemes are now in use in many states, including Alaska, Connecticut, Iowa, Maine, Massachusetts, Michigan, Minnesota, Nebraska, Nevada, New Jersey, New York, Oregon, Pennsylvania, Rhode Island, South Dakota, Washington, Wisconsin, and Wyoming. But there has been little formal analysis of the various schemes that are employed in these states and, as a result, the basis available for choice among them remains incomplete. This paper classifies the theoretical problems that must be resolved before a more careful comparison of these compulsory-arbitration schemes is possible, provides a brief overview of the work that has been done on each of these problems, and indicates what appear to be the most promising directions for future research along these lines.

Four kinds of compulsory arbitration are considered here: conventional compulsory arbitration (*CCA*), in which the arbitrator imposes a settlement of his (unrestricted) choice if negotiations break down; final-offer arbitration (*FOA*), in which the arbitrator must choose without compromise between bargainers' final offers if negotiations break down; multiple *FOA*, a variant of *FOA* originally suggested by Donn; and, on occasion, issue-by-issue *FOA*, which is like simple *FOA* except that the arbitrator is permitted to fashion his settlement from the components of bargainers' final offers. *CCA*, simple *FOA*, and issue-by-issue *FOA* are already in widespread use, while multiple *FOA*, which is similar but not identical to a scheme used in Eugene, Oregon, has been suggested by Donn and my 1979a article as an improvement on simple *FOA*.

In the literature of industrial labor relations, compulsory-arbitration schemes have been judged primarily by three criteria: the quality of the arbitral settlements they generate when negotiations break down; their freedom from bias, which is usually defined as the distortion of negotiated settlements away from what they would have been in ordinary bargaining, with both strikes and lockouts permitted; and the extent to which they create environments conducive to negotiated settlements.

An integrated analysis, in which bargainers choosing their stategies consider the effects of their actions on negotiated and arbitral settlements as well as on the probabilities of these possibilities, would be ideal. But in beginning the study of the effects of arbitration schemes, it is convenient, and probably not misleading, to simplify the problem by dividing it. Thus, I shall propose separate analyses of the quality of arbitral settlements, under noncooperative behavioral assumptions; the bias of negotiated settlements, under cooperative assumptions; and the probability of a negotiated settlement, under a blend of both noncooperative and cooperative assumptions. Each section of this paper in turn discusses existing work that is relevant to judging arbitration schemes by one of the above three criteria.

*University of California-San Diego. This research was supported by the National Science Foundation. Many of the observations made here evolved in discussions and correspondence with Clifford Donn, who by no means agrees with all, or even most, of them. I am grateful to him, and to Joel Sobel and participants in a workshop presentation at the University of Chicago, who also made helpful comments.

I. The Quality of Arbitral Settlements

This section is concerned with judging the quality of the settlements generated by the above-mentioned schemes when the arbitration mechanism must actually be used. The work discussed all assumes that bargainers expect an arbitral settlement (with certainty) and behave noncooperatively, given that expectation. My 1979a article contains the first formal analysis in this area. There, two main sets of results are obtained:

1) Suppose that bargainers are trying to negotiate a (possibly multidimensional) settlement, and that each bargainer knows both his opponent's current offer and how the arbitrator would choose between final offers if negotiations broke down. Then under reasonable assumptions, the game induced by *FOA* (or issue-by-issue *FOA*) has a unique, globally stable noncooperative equilibrium, at which both bargainers make their final offers $\bar{z}$, the settlement the arbitrator would impose in *CCA*; $\bar{z}$ then becomes the final settlement. These results do *not* require assumptions about the arbitrator's knowledge of bargainers' preferences or about bargainers' knowledge of each other's preferences.

2) A simple modification of *FOA*, called "multiple" *FOA*, yields equilibrium settlements that are at least as good as $\bar{z}$ for each bargainer, Pareto efficient (in terms of bargainers' preferences) when bargainers are well informed about each other's preferences, and Pareto superior to $\bar{z}$ even if bargainers are not well informed. Multiple *FOA* allows each bargainer to make two final offers, and requires the arbitrator to announce which bargainer has made the "best" offer. The other bargainer is then required to choose one of the first bargainer's offers.

These results have several interesting implications. *FOA* was originally proposed by Carl Stevens to counteract the common reluctance of bargainers whose negotiations are governed by *CCA* to make concessions, a reluctance usually explained by the observation that bargainers often appear to believe that if they fail to reach an agreement, the arbitrator's imposed settlement in *CCA* will tend to split the difference between their final offers. The results described above suggest strongly that the intent of the *FOA* statutes—to induce bargainers to make serious concessions and thereby reach their own agreements—is not likely to be realized by *FOA* in practice. If *FOA* can reconcile bargainers' differences, it can do so only by driving their offers toward what the arbitrator wants, leading ultimately to a settlement that is completely independent of bargainers' preferences, except as they affect the arbitrator's. Thus, the fairly high frequency of negotiated settlements under *FOA*, which is frequently presented in the industrial relations literature as evidence of its success, may indicate only that bargainers, while appearing to negotiate their own settlements, have correctly perceived the arbitrator's wishes and yielded to the incentives created by *FOA* to conform to them. This is in contrast to what obtains in *CCA*, which need not create the same incentives to yield to the arbitrator's wishes that *FOA* does. Multiple *FOA*, which is almost as simple to use as *FOA*, provides a partial remedy for these difficulties, preserving *FOA*'s ability to impose society's judgments about equity on the bargaining process—judgments inherent in any scheme that seeks to reconcile differences between bargainers who disagree about their "fair" shares of the pie—without creating incentives that prevent bargainers from reaching efficient settlements.

The major restrictive assumption that underlies the above results is that bargainers know the arbitrator's preferences. This makes little difference for *CCA*. But while this assumption is clearly a valid approximation for some applications of *FOA*, in which the arbitrator's discretion is severely limited by the arbitration statute or in which the arbitrator signals his intentions to bargainers, there are many other situations where it may be misleading. In fact, in much of the industrial relations literature, uncertainty about the arbitrator's intentions is thought to be an essential characteristic of *FOA*, which improves its performance (see, for example, Stevens, Donn, and Henry Farber), because it increases risk-averse bargainers' incentives to avoid an arbitral settlement. This view cannot be evaluated without a theory of the frequency of negotiated settlements, but the above results on arbitral settlements are quite likely to remain valid under uncertainty. *FOA*'s distinguishing feature is that it threatens bargainers with a settlement determined by the relative desirability to the arbitrator of their final offers. This creates incentives for bargainers to move their final offers closer to what they think the arbitrator wants, even if they are uncertain of his wishes. While it is harder to see where this process must terminate under uncertainty, *FOA* still reconciles differences mainly by encouraging bargainers to gear their concessions to the arbitrator rather than to each other, as efficiency requires.

Kalyan Chatterjee and Farber consider the effects of uncertainty about the arbitrator's preferences explicitly, working out examples in which bargaining is over a single issue. These suggest that *FOA* may lead to settlements that are less equitable *ex post* than those generated by *CCA*, a plausible conclusion. But *FOA* and *CCA* cannot really be compared on efficiency grounds in a one-dimensional framework, in which all feasible settlements are *ex post* Pareto efficient. Finally, Chatterjee and Farber's choice of the noncooperative equilibrium as a solution concept is a natural one in this context; and an equilibrium always exists in their examples, as in the certainty case discussed above. But existence is not guaranteed, even under quite strong assumptions, when bargainers in *FOA* are uncertain about the arbitrator's preferences. Thus, the proper choice of solution concept for *FOA* bargaining may require additional thought.

II. Bias and Negotiated Settlements

While compulsory-arbitration schemes have a direct effect on outcomes only in the event of impasse, the fact that they influence the bargaining environment means that, in general, they can be expected to exert an influence on negotiated settlements as well. This section discusses the problem of modeling that influence for the various schemes under consideration. Farber and Harry Katz and Farber study this influence for *CCA* and *FOA* in examples, assuming that negotiated settlements always give each bargainer a constant proportion, a measure of his "bargaining strength," of the difference between bargainers' certainty-equivalents of the prospect of an arbitral settlement. (This assumption is implicit in Farber, where effects on negotiated settlements are judged by shifts in the "contract zone"—the set of settlements that are at least as good for each bargainer as the prospect of an arbitral settlement.) The main conclusion of these analyses is that both *CCA* and *FOA* generate negotiated settlements that favor the less risk-averse bargainer, and that in *FOA* this bias is likely to be more severe than in *CCA*.

Two difficulties flaw this analysis, in addition to is possible dependence on the assumption that bargaining is one-dimensional and on the specific choice of functional forms. First, the treatment of bargaining strengths as fixed parameters, independent of bargainers' risk-aversion coefficients and of all other aspects of the bargaining problem, is difficult to justify in a model whose main focus is the effect of risk aversion. In game-theoretic and micro-economic theories of bargaining, there is generally an intimate relationship between risk aversion and bargaining strength. For example, in John Nash's 1950 theory, the greater a bargainer's risk aversion, the smaller his bargaining strength, *ceteris paribus*. One need not be a dogmatic believer in these theories to be skeptical of conclusions that rest on the assumption that bargaining strength is totally independent of risk preferences. The second criticism, which applies to Farber's analysis of *FOA* (but not to Farber and Katz's analysis of *CCA*, because in *CCA* the arbitral settlement is independent of bargainers' actions), is directed at Farber's assumption that the final settlement is derived from a contract zone determined by the final offers bargainers would make if they expected an arbitral settlement. For negotiated settlements, it is more natural to take the contract zone as determined by final offers directed at improving the negotiated settlement as much as possible, rather than at improving an arbitral settlement bargainers do not expect. Since an action that makes the arbitral settlement slightly worse for a bargainer but much worse for his opponent may increase the bargainer's utility at the negotiated settlement, these assumptions may lead to significantly different conclusions.

My 1979c paper models the effect of compulsory arbitration on negotiated settlements in a way that avoides the above criticisms, but at the cost of abstracting away from bargainers' possible uncertainty about the arbitrator's preferences. There, a simple model of Pareto-efficient bargaining, based on Nash's 1953 "variable-threats" model (developed for situations in which bargainers' actions can influence the disagreement outcome, as is true for all the schemes discussed here except *CCA*), is used to show that all four types of compulsory arbitration discussed here have precisely the same effect on negotiated settlements in the certainty case. While these schemes may have different biases in general, the differences are negligible, to a first approximation, unless there is significant uncertainty about the arbitrator's preferences.

If bargainers are uncertain about the arbitrator's or each other's preferences, formidable difficulties are encountered following this approach, because there is no widely accepted solution concept for variable-threats bargaining in such situations. But carrying the analysis as far as available tools allow might yield useful information.

III. The Probability of Impasse

It is a commonplace in the industrial relations literature that, for various reasons, negotiated settlements are to be preferred to the arbitral settlements that follow impasses. The effect of an arbitration scheme on the probability or frequency of impasse is, therefore, of great importance for judging such schemes. And the understanding of the influence of bargaining environments on the probability of impasse that is needed to judge this effect would yield large welfare gains in other areas. Nevertheless, there is almost no basis in the bargaining literature (in which it is generally assumed that efficient outcomes are always reached) for such judgments. Here, I shall criticize the approach to this problem that has been taken in the industrial relations literature, review some recent work that may ultimately lead to a firmer basis for analysis, and point out the shortcomings of that work from the standpoint of its applicability to the problem at hand.

A common implicit assumption in the industrial relations literature (see, for example, Farber, Farber-Katz, and Stevens) is that the probability that bargainers will negotiate their own settlement is determined by the size of the contract zone. Farber and Farber-Katz use this assumption to compare *CCA* and *FOA* in the examples described above, studying how the size and existence of the contract zone under these schemes is related to bargainers' risk preferences and their probabilistic beliefs about the arbitrator's preferences. They conclude that for both *CCA* and *FOA*, risk aversion on the part of bargainers and uncertainty about the arbitrator's preferences are the crucial factors that determine the size of the contract zone, and are necessary for it to exist at all; these conclusions are then used to make predictions about the probability of impasse. While their conclusions about the existence and size of the contract zone are questionable,[1] my main purpose here is to

[1] These conclusions are artifacts of Farber and Katz's assumptions that bargaining is one-dimensional and that bargainers cannot negotiate random settlements. Put simply, the existence of the contract zone depends on whether the expected utilities bargainers associate with the prospect of an arbitral settlement are in the utility-possibility set generated by the set of potential negotiated settlements. This is so whether bargainers are risk averters or risk lovers, and whether or not bargaining is one-dimensional. Suppose that (as Farber and Katz assume for much of their analysis) bargainers have identical priors about the arbitral settlement, and suppose further that random settlements are allowed (not because one expects to observe them, but because ruling them out needlessly complicates the analysis). Then there is *always* a contract zone, since bargainers' common distribution of possible arbitral outcomes is a potential negotiated settlement; in general, of course, they can do even better than this. When bargainers have different priors, the existence of the contract zone is guaranteed unless their priors are relatively too "optimistic"; in general, the size of the contract zone is

suggest that the view that the probability of impasse is directly determined by the size of the contract zone is a serious oversimplification, one which may be misleading. This is best accomplished by describing two theories that relate the frequency of impasse to the bargaining environment, theories which pay closer attention to bargainers' incentives, and may therefore yield more accurate predictions. These theories need to be developed further before they can be used to evaluate arbitration schemes other than *CCA*, because they assume that bargainers cannot affect the disagreement outcome; but I believe they are potentially quite useful, both in the design of arbitration schemes and in other, equally important, contexts.

My 1979b paper constructs a simple theory of bargaining impasses, building on Thomas Schelling's view of the bargaining process as a struggle between bargainers to commit themselves to favorable bargaining positions. Because bargaining impasses are generally inefficient, anything involving a positive probability of impasse is Pareto inefficient as well. In spite of this avoidable inefficiency, rational otherwise well-informed bargainers can respond to uncertainty about the outcome of the commitment process by taking actions that might lead to two successful incompatible commitments, with an impasse resulting from the irreversibility of commitment. In this model, changing the costs of disagreement in a way that expands the contract zone need not lower the equilibrium probability of impasse, in spite of the conventional wisdom to the contrary. Such changes alter the cost and benefits of commitment to various positions in a way that might lead bargainers to attempt commitment to positions that imply a higher probability of two successful incompatible commitments. But there is some theoretical evidence that the conventional wisdom is not completely invalid.

Chatterjee and William Samuelson have developed an interesting model of bargaining, which provides an alternative explanation of the occurrence of impasses. In their model, each bargainer makes an irreversible demand. If bargainers' demands are compatible, they split the difference between them; if they are incompatible, an impasse results. They then show that bargainers' equilibrium demands will generally involve "shading," or demanding more than their disagreement utilities. This leads to a positive probability that no bargain will be struck even when a contract zone exists. This model is also compatible with a "perverse" relationship between the size of the contract zone and the likelihood of impasse.

Before convincing applications to arbitration are possible, extensions in several directions—to variable-threats and multistage bargaining, among others—as well as more careful consideration of the choice of equilibrium concept, integration of the analyses described above, and examination of other possible explanations of disagreement are needed. But the central point is that to design better arbitration schemes, and better bargaining environments in general, impasses must be avoided as often as possible; to avoid impasses, one must understand what causes them. This seems to call for a partial reorientation of bargaining theory.

determined completely by how *ex ante* inefficient the prospect of disagreement is. Thus, the contract zone will tend to be large when bargainers are risk averse and uncertain about the arbitrator's likely actions, when they are risk lovers and too well informed about the arbitrator's actions, when arbitral settlements are disliked per se, when the arbitrator is poorly informed about bargainers' preferences, or when he seeks to promote goals other than their welfares alone (as he is frequently required by law to do). These last three points are obscured in Farber and Katz's models because there, all feasible settlements, and therefore the arbitral settlement, are *ex post* Pareto efficient, an extremely unlikely outcome in real bargaining.

REFERENCES

K. Chatterjee, "Comparison of Arbitration Procedures: Models with Complete and Incomplete Information," mimeo. 1980.

______ **and W. Samuelson,** "The Simple Economics of Bargaining," mimeo. 1979.

V. P. Crawford, (1979a) "On Compulsory-Arbitration Schemes," *J. Polit. Econ.*, Feb. 1979, *87*, 131–159.

______, (1979b) "A Theory of Disagreement

A CRITICAL LOOK AT RATIONAL EXPECTATIONS IN PRACTICE

Estimation and Control of Rational Expectations Models

By GREGORY C. CHOW*

Explanation and prediction of economic behavior are based mainly on the assumption that economic agents maximize. When the environment facing the economic agents is stochastic and the objective function is multiperiod, the techniques of optimal stochastic control can be employed to maximize the expectation of the objective function. Under the assumption of rational expectations, the economist assumes that his description (model) of the stochastic environment is identical with that of the economic agents (and with the "true" environment at least approximately) so that the expectations of economic variables as conceived by the economic agents are generated by the same stochastic model postulated by the economist. This paper explains how stochastic control techniques can be used to model economic behavior and to estimate the parameters of these models.

I. Model of Optimal Control and Its Estimation

Let the stochastic environment facing the economic agents and postulated by the economist be represented by the linear system

$$y_t = Ay_{t-1} + Cx_t + b + u_t \tag{1}$$

where y_t is a vector of state variables, x_t is a vector of variables subject to the control of the agents, and u_t is independent and identically distributed. As in my 1975 book (p. 153), a higher-order system including lagged variables $y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots$, etc. is converted into first-order where only y_{t-1} and x_t appear; the vector y_t incorporates x_t as a subvector by definition. Let the objective function which the agents maximize be quadratic

$$E_0 \sum_{t=1}^{T} (y_t - a_t)'K_t(y_t - a_t) \tag{2}$$

Then the optimal decision rule is linear in the state variables

$$x_t = G_t y_{t-1} + g_t \tag{3}$$

where (compare my 1975 book, pp. 178–79), G_t and g_t are obtained by solving the following equations backward in time from $t = T$ to $t = 1$:

$$(C'H_tC)G_t + C'H_tA = 0 \tag{4}$$

$$H_t - K_t - (A + CG_{t+1})'H_{t+1}(A + CG_{t+1}) = 0 \tag{5}$$

$$(C'H_tC)g_t + C'(H_tb_t - h_t) = 0 \tag{6}$$

$$h_t - K_ta_t - (A + CG_{t+1})'(h_{t+1} - H_{t+1}b_{t+1}) = 0 \tag{7}$$

with initial conditions $H_T = K_T$ and $h_T = K_Ta_T$.

In the above model, (1) may represent a macro-econometric model used by a central government whose objective is described by (2), with (3) representing the optimal rule for its conduct of monetary and fiscal policies. Alternatively, (1) may also represent the environment facing private economic agents (firms or consumers) whose objective function is (2); (3) then represent their optimal behavioral equations (demand functions for inputs, investment functions, or consumption functions). To explain economic behavior, the economist postulates (1) and (2), and assumes that the behavioral equations (3) result from maximization.

*Princeton University.

Assuming that the parameters of (2) satisfy $K_t = \beta^t K$ (β being a discount factor) and $a_t = \phi^t a$ (ϕ being a diagonal matrix), my 1980 article has provided statistical methods for the estimation of the parameters of (1) and (2) using time-series observations (y_t, x_t) when the coefficient matrix G_t in (3) becomes time invariant. The methods are maximum likelihood and two-stage least squares. One does not apply least squares to estimate G and g_t in (3) directly because they are derived from (or functions of) the parameters of (1) and (2), as given by (4)–(7). Thus, instead of estimating the parameters of the demand functions (3) for inputs directly, one estimates the parameters in (2) for the production function, etc. By the method of maximum likelihood, one uses equations (1) and (3) with a normal residual added to form a likelihood function. Noting that the parameters G and g_t of (3) are functions of the parameters of (1) and (2), one then maximizes the likelihood function with respect to the parameters of (1) and (2).

I present here a solution to the problem of maximum likelihood estimation of the parameters of (1) and (2) under the assumptions that $K_t = K$, $a_t = a$, that the system reaches a covariance stationary state, that the residual u_t in (1) is normal and serially uncorrelated, having a covariance matrix Σ, and that (3) contains an additive normal, serially uncorrelated residual which has a covariance matrix V and is uncorrelated with u_t.

Under the stated assumptions, the coefficients G and g in (3) are related to the parameters of (1) and (2) by equations (4)–(7), with all subscripts t omitted. The problem is the maximization of the likelihood function subject to these four constraints. I form a Lagrangian expression which combines the *log*-likelihood with these constraints

$$L = \text{constant} - \frac{n}{2} log|\Sigma| - \frac{n}{2} log|V|$$

$$-\tfrac{1}{2} tr\left[\Sigma^{-1}(Y' - AY'_{-1} - CX' - bz')\right.$$

$$\left.\times (Y - Y_{-1}A' - XC' - zb')\right]$$

$$-\tfrac{1}{2} tr\left[V^{-1}(X' - GY'_{-1} - gz')\right.$$

$$\left.\times (X - Y_{-1}G' - zg')\right] - tr[\Omega(4)]$$

$$-\tfrac{1}{2} tr[\Phi(5)] - \omega'[(6)] - \phi'[(7)]$$

$$-\tfrac{1}{2}\theta[tr(KK) - r]$$

where Y is an $n \times p$ matrix of observations on the endogenous variables; Y_{-1} is an $n \times p$ matrix of observations on the lagged endogenous variables; X is an $n \times q$ matrix of observations on the control variables; z represents a dummy variable being a vector consisting of n ones; $\Omega(p \times q)$ and $\Phi = \Phi(p \times p)$ are matrices of Lagrangian multipliers; $\omega(q \times 1)$ and $\phi(p \times 1)$ are vectors of Lagrangian multipliers; the numbers 4–7 in parentheses denote the corresponding constraints (with all subscripts t omitted); and the last constraint $tr(KK) = r$ serves to normalize the matrix K, r being the number of target variables, or the number of nonzero diagonal elements in K. The unknowns in this problem consist of Σ, V, A, C, b, G, H, K, g, h, and a. In my 1980 article and 1981 book, I show how to maximize L with respect to these parameters. Thus the problem of estimating this rational expectations model is solved.

II. Behavior of Private Agents Given a Government Decision Rule

Let there be two sets of decision makers (players) whose control variables are x_{1t} and x_{2t}, respectively. The model (1) becomes

$$(8) \quad y_t = Ay_{t-1} + C_1x_{1t} + C_2x_{2t} + b + u_t$$

Let the objective function of the ith player be, with time invariant K_i and a_i,

$$(9) \quad E_0 \sum_{t=1}^{T} (y_t - a_i)'K_i(y_t - a_i) \quad (i = 1,2)$$

Further, let their decision rules (however arrived at) be written as

$$(10) \qquad x_{it} = G_iy_{t-1} + g_i \quad (i = 1,2)$$

If player 2 represents the government which follows a fixed decision rule for x_{2t}, the environment facing the private economic agents (player 1) will be

$$(11) \quad y_t = (A + C_2G_2)y_{t-1} + C_1x_{1t} + (b + C_2g_2) + u_t \equiv A_1y_{t-1} + C_1x_{1t} + b_1 + u_t$$

Assuming that the private economic agents facing the environment (11) solve their optimal control problem to obtain their behavioral equation $x_{1t} = G_1y_{t-1} + g_1$, one concludes that their behavioral equation will change as the government's decision rule changes. Therefore, an econometrician cannot rely on stable behavioral equations for the private sector to evaluate the consequences of changing government policies, as Robert Lucas has stressed. On the other hand, if the econometrician has estimated the parameters of (1) or (8) and (2) for the private agents, he can infer their behavior (3) once the government's decision rule is known, by solving the optimal control problem with a new environment (11).

Specifically, let the government adhere to a feedback control policy $x_{2t} = G_2y_{t-1} + g_2$, player 1 will face (11) as its environment and adopt the optimal equilibrium strategy $x_{1t} = G_1y_{t-1} + g_1$ where, on account of (4)–(7),

$$(12) \quad C_1'H_1C_1G_1 + C_1'H_1(A + C_2G_2) = 0$$

$$(13) \quad H_1 - K_1 - (A + C_2G_2 + C_1G_1)'H_1 (A + C_2G_2 + C_1G_1) = 0$$

$$(14) \quad C_1'H_1C_1g_1 + C_1'[H_1(b + C_2g_2) - h_1] = 0$$

$$(15) \quad [I - (A + C_2G_2 + C_1G_1)']h_1 - K_1a_1 - (A + C_2G_2 + C_1G_1)'H_1(b + C_2g_2) = 0$$

Given G_2, equations (12) and (13) can be solved to obtain G_1 and H_1. Given g_2 in addition, equations (14) and (15) can be solved to obtain g_1 and h_1. Equations (12)–(15) show exactly how the behavior (G_1, g_1) of the private sector will change as the government policy rule (G_2, g_2) changes. An interesting question, raised by James Tobin to the author, is under what circumstances the response of the private sector to a new government policy rule will not lead to changes in the behavior of the economic system, thus making government policy ineffective. I cannot provide an answer here, except to point out that it depends partly on the number of instruments available to the private sector.

III. Optimal Control Rule for the Government

Once the government can predict how its decision rule would affect the behavior of the private economic agents, it can choose a rule to maximize its objective function, assuming all parameters of (8) and (9) to be known. An algorithm for the government is given in my 1981 book (ch. 17). The problem is to find the optimal strategy of the dominant player in a two-person dynamic game. I will derive a pair of optimal steady-state strategies (G_1, g_1) and (G_2, g_2) for the two players when the system is in a covariance-stationary equilibrium, assuming that a_{1t}, K_{1t}, a_{2t}, and K_{2t} are all time invariant.

Given any government policy (G_2, g_2) and the associated strategy of the private sector as given by equations (12)–(15), the behavior of the system (8) is determined. In a covariance-stationary equilibrium, the system will have a mean vector $\bar{y}$ and a covariance matrix $\Gamma = E(y_t - \bar{y})(y_t - \bar{y})'$ which satisfy (see my 1975 book, pp. 51–52):

$$(16) \quad (I - A - C_1G_1 - C_2G_2)\bar{y} - b - C_1g_1 - C_2g_2 = 0$$

$$(17) \quad \Gamma - (A + C_1G_1 + C_2G_2)\Gamma (A + C_1G_1 + C_2G_2)' - Eu_tu_t' = 0$$

The problem of the government, here treated as the dominant player, is to minimize its loss function

$$\tfrac{1}{2}E(y_t-a_2)'K_2(y_t-a_2)$$
$$=\tfrac{1}{2}tr(K_2\Gamma)+\tfrac{1}{2}(\bar{y}-a_2)'K_2(\bar{y}-a_2)$$

with respect to G_2 and g_2 in its feedback control equation, subject to the constraints (12)–(17). This problem can be solved by forming the Lagrangian expression

$$L=\tfrac{1}{2}tr(K_2\Gamma)+\tfrac{1}{2}(\bar{y}-a_2)'K_2(\bar{y}-a_2)$$
$$-\omega'(14)-\phi'(15)-\lambda'(16)$$
$$-tr\{\Omega(12)\}-\tfrac{1}{2}tr\{\Phi(13)\}-\tfrac{1}{2}tr\{\Psi(17)\}$$

where ω, ϕ, λ, Ω, $\Phi=\Phi'$, and $\Psi=\Psi'$ are vectors and matrices of Lagrangian multipliers and, for brevity, the equation number in parentheses denotes the corresponding constraint.

Using the differentiation rule $\partial\, tr(AB)/\partial A=B'$, I obtain the following equations, with R denoting $A+C_1G_1+C_2G_2$,

$$\text{(18)}\qquad \frac{\partial L}{\partial g_1}=-C_1'H_1C_1\omega+C_1'\lambda=0$$

$$\text{(19)}\qquad \frac{\partial L}{\partial h_1}=C_1\omega-(I-R)\phi=0$$

(20)
$$\frac{\partial L}{\partial g_2}=-C_2'H_1C_1\omega+C_2'H_1R\phi+C_2'\lambda=0$$

$$\text{(21)}\qquad \frac{\partial L}{\partial \bar{y}}=K_2(\bar{y}-a_2)-(I-R')\lambda=0$$

$$\text{(22)}\qquad \frac{\partial L}{\partial G_1}=C_1'[\,H_1C_1\Omega'+H_1R\Phi+\Psi R\Gamma$$
$$+h_1\phi'+H_1(b+C_2g_2)\phi'+\lambda\bar{y}'\,]=0$$

$$\text{(23)}\qquad \frac{\partial L}{\partial G_2}=C_2'[\,H_1C_1\Omega'+H_1R\Phi+\Psi R\Gamma$$
$$+h_1\phi'+H_1(b+C_2g_2)\phi'+\lambda\bar{y}'\,]=0$$

$$\text{(24)}\qquad \frac{\partial L}{\partial H_1}=-\Phi+R\phi R'-C_1g_1\omega'C_1'$$
$$-C_1\omega g_1'C_1'-(b+C_2g_2)\omega'C_1'$$
$$-C_1\omega(b+C_2g_2)'+(b+C_2g_2)\phi'R$$
$$+R'\phi(b+C_2g_2)'=0$$

$$\text{(25)}\qquad \frac{\partial L}{\partial \Gamma}=K_2-\Psi+R'\Psi R=0$$

To solve these equations, let us start with some tentative solution for G_2, G_1, and H_1. Equations (18), (19), (20), and (21) imply, respectively, with $P_1=C_1(C_1'H_1C_1)^{-1}C_1'$,

$$\text{(18')}\qquad \omega=(C_1'H_1C_1)^{-1}C_1'\lambda$$

$$\text{(19')}\qquad \phi=(I-R)^{-1}P_1\lambda$$

$$\text{(20')}\qquad C_2'\{I-H_1[\,I-R(I-R)^{-1}\,]P_1\}\lambda=0$$

$$\text{(21')}\qquad \lambda=(I-R')^{-1}K_2(\bar{y}-a_2)$$

Equations (21) and (16) give

$$\text{(26)}\qquad \lambda=(I-R')^{-1}K_2[\,(I-R)^{-1}$$
$$\times(b+C_1g_1+C_2g_2)-a_2\,]$$

Combining (26) with (20′), I get

(27)
$$C_2'\{I-H_1[\,I-R(I-R)^{-1}\,]P_1\}(I-R')^{-1}$$
$$K_2[\,(I-R)^{-1}(b+C_1g_1+C_2g_2)-a_2\,]=0$$

With G_2, G_1, and H_1 given, equations (27), (14), and (15) can be solved for g_2, g_1, and h_1. Equations (26), (18′), and (19′) are then used to find λ, ω, and ϕ, while equation (16) is used to compute $\bar{y}$.

I now solve equations (22)–(25). Equation (24) is used to solve for Φ iteratively, that is $\Phi^{(i+1)}=R\Phi^{(i)}R'+$known matrix. Equations (22) and (23) imply

(22′)
$$\Omega'=-(C_1'H_1C_1)^{-1}C_1'[\,H_1R\Phi+\Psi R\Gamma+\ldots\,]$$

$$(23')\quad C_2'[I-H_1P_1][H_1R\Phi+\Psi R\Gamma+h_1\phi' + H_1(b+C_2g_2)\phi'+\lambda\bar{y}']=0$$

Since (17) and (25) can be used to compute Γ and Ψ, respectively, (23′) after being postmultiplied by Γ^{-1} can be solved for G_2 iteratively, i.e.,

$$C_2'[I-H_1P_1]\Psi C_2G_2=C_2'[I-H_1P_1] \times[H_1R\Phi\Gamma^{-1}+\Psi(A+C_1G_1)+\dots]$$

where I have recalled $R=(A+C_1G_1+C_2G_2)$. Having thus obtained a new matrix G_2, let us continue with the iterative process by returning to the beginning of the preceding paragraph.

Mathematically, the solution to the two-person dynamic game formulated above under a Nash (or Cournot) equilibrium is simpler, for each player would treat the other's strategy as given, without being affected by his own strategy. Given (G_2, g_2), player 1 would find (G_1, g_1) by equations (12)–(15) as before. Symmetrically, given (G_1, g_1), player 2 would find (G_2, g_2) by solving an identical set of equations with subscripts 1 and 2 interchanged. A Nash equilibrium is found by solving these two sets of equations.

IV. Estimating Models of Dynamic Game

When x_{2t} in equation (8) represents the policy instruments of the government and the government is treated as the dominant player, I will study the estimation problem in two stages. First, assuming that the government adheres to a policy rule $x_{2t}=G_2y_{t-1}+g_{2t}$, which is decided upon by whatever means, I will consider the estimation of the parameters of (9) and (10) for $i=1$ under the assumption that the private sector behaves optimally. Second, from the above framework, I take the next step by assuming that the government is also trying to maximize (9) for $i=2$ and consider the estimation of the parameters of its objective function as well.

For the first problem, the stochastic environment facing the private sector is given by (11). The methods of my 1980 article can be applied to estimate the parameters of (11) and (10) for $i=1$. In order to solve the second and more difficult problem of estimating K_{2t} and a_{2t}, I treat a more restrictive case by introducing the assumption that K_{1t}, a_{1t}, K_{2t}, and a_{2t} are all time invariant. Given K_1, a_1, K_2, a_2, and the parameters of (4), we can apply the method of Section III to find (G_1, g_1) and (G_2, g_2); thus the likelihood function can be evaluated. A gradient method can in principle be applied to maximize the likelihood with respect to these parameters, but this numerical maximization problem requires further investigation.

The estimation problem for a dynamic game model under a Nash equilibrium is simpler. I can apply iterative techniques by considering this estimation problem in two stages. First, assuming tentatively that the government adheres to a policy rule (G_2, g_{2t}), I will maximize the likelihood function by the method of my 1980 article with respect to the parameters of (8) and $K_{1t}=\beta_1^tK_{10}$ and $a_{1t}=\phi_1^ta_{10}$ under the assumption that the private sector behaves optimally. This estimation procedure assumes optimal behavior (G_1, g_{1t}) of the private sector, with (G_2, g_{2t}) taken as given. Second, assuming that the private sector adheres to the policy (G_2, g_{2t}) as determined above, I find maximum likelihood estimates of the parameters of (8) and $K_{2t}=\beta_2^tK_{20}$ and $a_{2t}=\phi_2^ta_{20}$ under the assumption that the government behaves optimally. Similarly, this estimation procedure assumes optimal behavior (G_2, g_{2t}) of the government, with (G_1, g_{1t}) taken as given. I now go back to step one, and iterate back and forth until convergence.

V. Further Applications

The above dynamic game model can be applied to other economic problems, such as modelling the behavior of two government decision makers, including the executive branch and the Federal Reserve in the

United States for example, or two private economic sectors or agents, as in the dynamic theory of oligopoly. As pointed out in my 1981 book (ch. 13), it can also be applied to model a planned economy such as the Soviet Union where the government is assumed to determine its behavior (3) by maximizing an objective function (2), given the environment (1). If an econometric model for the planned economy is constructed without postulating optimizing behavior by the government, it may consist of equations (1) and (3), the latter including the government's investment functions. If the investment and other functions of the government are assumed to be derived by the maximization of (2), an econometrician may estimate the parameters of (1) and (2) by the method suggested in Section I, rather than the parameters of (1) and (3) directly. For the purpose of forecasting the future, (1) and (2) will be used, while the behavior described by (3) will be derived by maximization. This method can allow for possible changes in government priorities and/or the environment (1).

A major assumption made to solve the estimation problems of this paper is that during the sample period the policy rules of the players remained the same. This assumption is strong, but not any stronger than the assumption necessary for the estimation of structural equations in simultaneous equations models. For the latter problem, knowledge of the structure is required to predict how the reduced-form corresponding to our equation (3) will change in response to policy changes, but to estimate the structure one ordinarily assumes constant parameters during the sample period.

REFERENCES

Gregory C. Chow, *Analysis and Control of Dynamic Economic Systems*, New York 1975.

______, "Estimation of Rational Expectations Models," *J. Econ. Dynamics, Control*, Aug. 1980, *2*, 241–55.

______, *Econometric Analysis by Control Methods*, New York 1981.

R. Lucas, Jr., "Econometric Policy Evaluation: A Critique," *J. Monet. Econ.*, Suppl., 1976, *1*, 19–46.

Thomas J. Sargent, *Macroeconomic Theory*, New York 1979.

Capital Mobility and Devaluation in an Optimizing Model with Rational Expectations

By MAURICE OBSTFELD*

This paper examines the effects of exchange-rate policies when individuals maximize lifetime utility on the basis of rational expectations about the future. The economy studied is one in which the authorities allow free mobility of capital under a crawling-peg exchange-rate regime. Many industrializing economies have adopted a crawling peg as a means of reconciling disparate inflation rates at home and abroad, and some recent efforts to use the rate of crawl as an instrument of anti-inflation policy have attracted considerable interest (see Carlos Díaz Alejandro). Tools similar to those employed here have been applied by Guillermo Calvo (forthcoming) to study this type of exchange-rate management under conditions of capital immobility.

An advantage of the explicit optimization framework is the light it throws on the interaction between private economic decisions and the balance sheets of the government and, particularly, the central bank. Although the literature on open-economy financial policy has largely ignored such considerations—the exceptions include papers by Alan Stockman and myself (1980a; forthcoming)—the extent to which the public internalizes official asset holdings has obvious consequences for the efficacy of official intervention in asset markets. An implication of the model explored below is that the stock of central bank foreign reserves, when allowed to earn interest, will be perceived by the public as part of its own wealth. In this setting, devaluation loses its short-run real effects, for a discrete rise in the price of foreign exchange occasions a proportional issue of central bank money, a transfer of interest-bearing foreign bonds from the public to the bank, and nothing more.

I. The Model

The economy I consider is inhabited by a representative household, which derives utility from consuming a single composite commodity and holding real money balances. Household wealth is divided between high-powered money holdings (there is no banking system) and holdings of an internationally traded bond having a fixed foreign currency face value. The economy is small, and therefore can influence neither the (positive) world bond rate ρ, nor the foreign currency price of the consumption good P^*, both of which are taken to be fixed.[1] The domestic currency price of consumption P is linked to the world price by the relationship $P=EP^*$, where E, the exchange rate, is the price of foreign money in terms of domestic money.

The central bank causes the exchange rate of depreciate according to a pre-announced schedule by standing ready at each instant to trade home currency for foreign exchange on the pre-announced terms. This type of exchange-rate regime has been studied, in somewhat different contexts, by Calvo (1979; forthcoming), Carlos Rodríguez, and others. It is assumed that the official schedule calls for a constant, nonnegative rate of devaluation, ε, which must equal the domestic inflation rate $\dot{P}/P$ at all times.

The representative household's instantaneous utility function is written as $u_t = u(c_t, m_t)$, where c_t is the family's consumption rate and m_t represents its nominal

*Assistant professor, Columbia University. I am grateful for conversations with Guillermo Calvo, and for very helpful comments from Dale Henderson and John Taylor. Financial support from the Columbia University Council for the Social Sciences is acknowledged with thanks.

[1] The assumption that P^* is constant implies that bonds are, in effect, indexed to the consumption good.

money holdings M_t, deflated by P_t. The function $u(\cdot,\cdot)$ is positive, twice continuously differentiable, increasing in both its arguments and strictly concave. Both consumption and money services are normal goods.

Maximization of the functional

$$(1) \qquad \int_0^{\infty} u_t e^{-\Delta_t} dt$$

is the household's lifetime objective. The discount factor Δ_t is defined as

$$(2) \qquad \Delta_t \equiv \int_0^t \delta_s ds$$

where δ_s is the household's instantaneous subjective discount rate at time s. To enable the economy to attain a stationary state at the constant world interest rate ρ, I will adopt the approach of Hirofumi Uzawa, in which the instantaneous discount rate at time s is a positive function of the contemporaneous utility level, $\delta_s = \delta(u_s)$. It is convenient to endow $\delta(\cdot)$ with the properties $\delta'(u)>0$, $\delta''(u)>0$, and $\delta(u)-u\delta'(u)>0$, for all u. These properties facilitate solution of the household's maximization problem, but are in no way intrinsic to the notion of an endogenous rate of time preference.[2]

In maximizing (1), the family is bound by three constraints. The first is a (stock) wealth constraint, which states that marketable real assets at time t, a_t, must equal the sum of real bond holdings b_t and real balances:

$$(3) \qquad a_t = b_t + m_t$$

The second is a (flow) savings constraint,

$$(4) \qquad \dot{a}_t = y + \rho b_t + \tau_t - c_t - \varepsilon m_t$$

where y is the economy's (fixed) output, τ_t net real transfers from the government, and εm_t the inflation tax on cash balances. The final constraint is the intertemporal budget constraint,

$$(5) \qquad \int_0^{\infty} (c_t + (\varepsilon+\rho)m_t - \tau_t) e^{-\rho t} dt \leqslant y/\rho + b_0 + m_0$$

The consolidated budget constraint of the government and central bank plays a key role in the developments below. It embodies the very important assumption that central bank reserves earn interest at the world rate, ρ. The government makes net transfer payments to domestic residents, but does not consume goods. In the absence of interest-bearing government debt, any excess of transfer payments over central bank foreign interest earnings must be financed by domestic credit creation. I assume that the level of government transfers is continuously varied in such a way that the rate of real domestic credit expansion equals 100ε percent of the stock of real balances at each moment. This implies a public-sector budget constraint of the form $\tau_t - \rho r_t = \varepsilon m_t$, where r_t denotes the central bank's real reserve holdings at time t.

II. The Perfect-Foresight Equilibrium Path

For any expected path $\{\tau_t\}$ of transfer payments, maximization of (1) subject to the constraints (3)–(5) yields the household's preferred paths for consumption ($\{c_t^*\}$), real balances ($\{m_t^*\}$), and external claims ($\{b_t^*\}$), together with an implied path for central bank reserves ($\{r_t^*\}$). The economy's perfect-foresight equilibrium path has the property that $\{\tau_t\}$, $\{m_t^*\}$ and $\{r_t^*\}$ are mutually consistent, in the sense that the government's budget constraint $\tau_t = \rho r_t^* + \varepsilon m_t^*$ is satisfied at each instant.

The first step in finding this path is to derive necessary conditions for an optimal household plan, contingent on an assumed path $\{\tau_t\}$ of transfers. (See my 1980 papers.) Using (2), one can simplify the maximization problem by a change of variables from t to Δ in (1) and (4) (see Uzawa). Letting $\lambda = \lambda_t$ denote the costate variable, the shadow price of wealth in utility terms at

[2] My 1980 papers contain a more complete discussion of the conditions on $\delta(\cdot)$.

time t, necessary conditions for an optimal program are

$$(6)\quad \lambda=\frac{u_c(\delta-\delta' u)}{\delta+\delta' u_c(y+\rho b+\tau-c-\varepsilon m)}$$

$$(7)\quad u_m/u_c=\varepsilon+\rho$$

$$(8)\quad \dot{\lambda}=\delta(u)\frac{d\lambda}{d\Delta}=\lambda(\delta(u)-\rho)$$

and the flow constraint (4). Given $\varepsilon+\rho$, (7) defines real money demand as an implicit function ϕ of consumption, with $\phi'(c)>0$.

The perfect-foresight assumption is imposed by adding to (4) and (6)–(8) the requirement that anticipated transfers $\{\tau_t\}$ and actual transfers $\{\rho r_t+\varepsilon m_t\}$ coincide. The resulting system of differential equations satisfies both the conditions necessary for optimality and the consistency condition linking expected transfers and the government budget constraint. Using the definition of ϕ, the system is described by the equations

$$\begin{aligned}(9)\quad \lambda=\big(&u_c(c,\phi(c))\{\delta(u(c,\phi(c)))\\ &-\delta'(u(c,\phi(c)))u(c,\phi(c))\}\big)\\ &\div\big(\delta(u(c,\phi(c)))+\delta'(u(c,\phi(c)))\\ &\times u_c(c,\phi(c))(y+\rho f-c)\big)\end{aligned}$$

$$(10)\quad \dot{\lambda}=\lambda(\delta(u(c,\phi(c)))-\rho)$$

$$(11)\quad \dot{f}=y+\rho f-c$$

where $f\equiv b+r$ denotes the stock of claims on foreigners owned by the country as a whole. Equation (11) displays the equilibrium rate of external asset accumulation as the difference between national income and absorption. It is derived by noting that $\dot{a}=\dot{b}+\dot{m}$, while $\dot{m}$ must equal $\dot{r}$ because domestic credit expansion just compensates money-holders for the real depreciation of their cash balances.[3]

The stationary-state levels of consumption and foreign claims are those such that $\dot{\lambda}=\dot{f}=0$. By writing (9)–(11) as a system in c and f, and then taking its linear approximation in a neighborhood of this stationary state, one obtains a graphical representation of the perfect-foresight equilibrium path. Equation (9) implies a relationship of the form $c=c(\lambda,f)$; thus,

$$(12)\quad \dot{c}=c_\lambda\dot{\lambda}+c_f\dot{f}\equiv\theta(c,f)$$

where (9)–(11) have been used to eliminate $\dot{\lambda}$, λ, and $\dot{f}$ in defining $\theta\,(\cdot,\cdot)$. The local linearization of (12) is

$$\begin{aligned}(13)\quad \dot{c}\simeq{}&\frac{\bar{u}_c(\rho-\bar{\delta}'\bar{u})\,\bar{\delta}'\bar{u}_m\bar{\phi}'}{\Delta}(c-\bar{c})\\ &+\frac{\bar{u}_c^2\,\bar{\delta}'(\rho-\bar{\delta}'\bar{u})\rho}{\Delta}(f-\bar{f})\end{aligned}$$

where a bar indicates a stationary-state value, and

$$\begin{aligned}\Delta={}&(\rho-\bar{\delta}'\bar{u})\big(\bar{u}_{cc}+\bar{u}_{cm}\bar{\phi}'-\bar{u}_c(\bar{\delta}'/\rho)\bar{u}_m\bar{\phi}'\big)\\ &-\bar{u}_c\bar{u}\bar{\delta}''(\bar{u}_c+\bar{u}_m\bar{\phi}')<0\end{aligned}$$

Figure 1 displays the phase portrait of the system described by (11) and (13). The locus of points along which $\dot{f}=0$ is upward sloping, with foreign claims increasing to its right and decreasing to its left. The locus along which $\dot{c}=0$ is negatively sloped; consumption is falling above this schedule and rising below it (for $\bar{\theta}_c<0$).

The stationary state $(\bar{c},\bar{f})$ is a saddlepoint: for an initial value f_0 of the predetermined variable, there is a unique initial consumption level c_0 placing the economy on a convergent path. Any consumption

[3] The time derivative of the nominal money supply is just the sum of the time derivatives of its foreign and domestic source components, $M_t=P_t\dot{r}_t+\varepsilon M_t$. It follows that $\dot{m}_t=M_t/P_t-\varepsilon m_t=\dot{r}_t$.

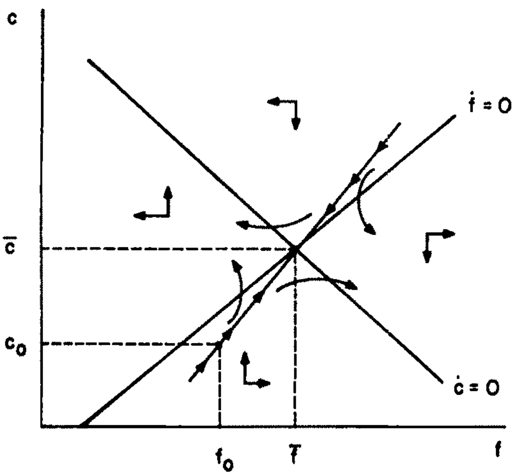


FIGURE 1

level exceeding c_0 initiates a trajectory along which the intertemporal budget constraint (5) is violated. Such paths are infeasible, and may thus be ruled out. Paths initiated by consumption levels below c_0 are feasible, but are not optimal from the household's standpoint, given the associated paths of expected future transfer payments. The convergent trajectory can be shown to be optimal as well as feasible. It must, therefore, be the economy's perfect-foresight equilibrium path. For a given initial stock of external claims, the equilibrium consumption level is unambiguously determined.

III. Exchange-Rate Policies

To illustrate the workings of the model, I consider in this section two types of exchange-rate policy, a one-time unanticipated devaluation of the currency (a discrete increase in E) and a permanent unanticipated increase in the *rate* of devaluation, ε. These are the two policies compared by Calvo (forthcoming) for a utility-maximizing economy in which domestic money is the only privately owned asset.

An unanticipated devaluation leaves the two schedules in Figure 1 unchanged, and has no effect on the national stock of foreign claims, which can change only over time. The devaluation thus leaves the economy unperturbed: it has no real consequences. Just as in the traditional one-asset framework (see Rudiger Dornbusch and Calvo, forthcoming), the devaluation brings about a sharp rise in the price level and fall in real balances. But in a setting of capital mobility, the fall in real balances is only momentary. An incipient excess demand for money exerts downward pressure on the exchange rate, and this forces the central bank to intervene in the asset market, purchasing foreign bonds and issuing money until the public's real balances have been restored to their initial level.

It is important to realize why this transfer of bonds from the public to the central bank reverses the wealth effect on consumption typically associated with the fall in the real value of privately held, marketable assets (see, for example, Jacob Frenkel and Rodríguez). Reserves acquired by the bank continue to earn interest that must be returned to the public in the form of higher transfer payments. The public, in turn, anticipates and capitalizes this income stream, and so, perceives no change in its *overall* wealth when real balances are again at their initial level. Accordingly, the devaluation does not alter consumption. While there is an instantaneous increase in reserves, there ensues no flow surplus in the balance of payments or current account.[4]

In contrast, an unanticipated increase in the *rate* of devaluation does have real consequences, for it shifts the $\dot{c}=0$ locus. By (8) the economy's stationary utility level $\bar{u}$ brings its rate of time preference into equality with the world interest rate, and is thus independent of the policy parameter ε. The increase in ε raises the opportunity cost of holding real balances, inducing a long-run substitution of consumption for real balances along the utility contour $\bar{u}$. This means that the $\dot{c}=0$ locus must shift to the right. In the new stationary state, a higher consump-

[4] These conclusions naturally require modification when reserves bear no interest, when some international lending is denominated in domestic currency, or when economic units have finite lives and leave no bequests.

tion level is financed by a higher stock of interest-bearing foreign claims. Money is not superneutral under the present assumptions.

What are the characteristics of the transition path? The saddlepath relevant after the increase in ε passes below the original long-run equilibrium, $(\bar{c}, \bar{f})$. If the economy is initially at rest, there is a sharp fall in consumption as the new optimal trajectory is attained. Together with the fall in the real return on money, this occasions a drop in real money demand, accommodated by central bank sales of foreign-exchange reserves to the public.

The initial postdisturbance equilibrium is a position of current-account surplus; consumption and the national stock of foreign assets grow along the path to the new long-run equilibrium. Necessary condition (7) shows that desired real balances and so, reserves, must grow as well during the adjustment process. Ultimately, however, the reserves gained during the transition do not make up for those lost during the initial portfolio adjustment in reaction to the change in ε. A higher asymptotic level of reserves would be inconsistent with the requirement that real balances be lower in the new stationary state than in the original one.

This time path of reserves and consumption stands in interesting contrast to the one found by Calvo (forthcoming) in a context of capital immobility. When there is only one asset, there can be no portfolio shift in response to an increase in the rate of depreciation, and so a discrete fall in real balances is precluded. Rather than overshooting their eventual, lower levels, real money and reserves decline monotonically over time. From an initial position of long-run equilibrium, consumption must rise on impact to induce the implied current-account deficit.

REFERENCES

G. A. Calvo, "Devaluation: Levels vs. Rates," *J. Int. Econ.*, forthcoming.

______, "An Essay on the Managed Float —The Small Country Case," unpublished manuscript, Columbia Univ. 1979.

C. F. Díaz Alejandro, "Southern Cone Stabilization Plans," unpublished manuscript, Yale Univ. 1979.

R. Dornbusch, "Real and Monetary Aspects of the Effects of Exchange Rate Changes," in R. Z. Aliber, ed., *National Monetary Policies and the International Financial System*, Chicago 1974.

J. A. Frenkel and C. A. Rodríguez, "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

M. Obstfeld, "The Capitalization of Income Streams and the Effects of Open-Market Policy under Fixed Exchange Rates," *J. Monet. Econ.*, forthcoming.

______, (1980a) "Macroeconomic Policy, Exchange-Rate Dynamics, and Optimal Asset Accumulation," unpublished manuscript, Columbia Univ. 1980.

______, (1980b) "Aggregate Spending and the Terms of Trade: Is there a Laursen-Metzler Effect?," unpublished manuscript, Columbia Univ. 1980.

C. A. Rodríguez, "Managed Float: An Evaluation of Alternative Rules in the Presence of Speculative Capital Flows," *Amer. Econ. Rev.*, Mar. 1981, *71*, 256–60.

A. C. Stockman, "Monetary Control and Sterilization under Pegged Exchange Rates," unpublished manuscript, Univ. Rochester 1979.

H. Uzawa, "Time Preference, the Consumption Function, and Optimum Asset Holdings," in J. N. Wolfe, ed., *Value, Capital, and Growth: Papers in Honor of Sir John Hicks*, Chicago 1968.

The Determinants of the Variability of Stock Market Prices

By Sanford J. Grossman and Robert J. Shiller*

The most familiar interpretation for the large and unpredictable swings that characterize common stock price indices is that price changes represent the efficient discounting of "new information." It is remarkable given the popularity of this interpretation that it has never been established what this information is about. Recent work by Shiller, and Stephen LeRoy and Richard Porter, has shown evidence that the variability of stock price indices cannot be accounted for by information regarding future dividends since dividends just do not seem to vary enough to justify the price movement. These studies assume a constant discount factor. In this paper, we consider whether the variability of stock prices can be attributed to information regarding discount factors (i.e., real interest rates), which are in turn related to current and future levels of economic activity.

The appropriate discount factor to be applied to dividends which are received k years from today is the marginal rate of substitution between consumption today and consumption k periods from today. We use historical data on per capita consumption from 1890–1979 to estimate the realized value of these marginal rates of substitution. Theoretically, as LeRoy and C. J. La Civita have also noted independently of us, consumption variability may induce stock price variability whose magnitude depends on the degree of risk aversion.

Robert Hall also studied these marginal rates of substitution and concluded that consumption is a random walk. We show that if current consumption and dividends are the best predictors of future consumption and dividends in Hall's sense, then the discount factor applied to stock prices would not vary. The variability of stock prices implies they do vary, so we conclude that consumers must have a better method for forecasting future consumption than using only current consumption (for example, consumers may know when the economy is in a recession).

I. Stock Returns and the Marginal Rate of Substitution

Consider a consumer who can freely buy or sell asset i and whose utility can be written as the present discounted value of utilities of consumption in future years $U_t = \sum_{k=0}^{\infty}\beta^k u(C_{t+k})$, where $\beta = 1/(1+r)$ and r is the subjective rate of time preference. A necessary condition for his holdings of the asset at t to be optimal, given that the consumer maximizes the expectation at time t, of this utility function is

$$(1)\qquad u'(C_t)P_{it} = \beta E[\,u'(C_{t+1})(P_{it+1} + D_{it+1})|I_t]$$

where P_{it} is the real price (in terms of the single consumption good or "market basket" C_t) of asset i and D_{it+1} is the real dividend paid at $t+1$ to holders of record at t. The term E denotes mathematical expectation, conditional here on I_t which is all the information about the future which the agent possesses at time t. The left-hand side of (1) is the cost in terms of foregone current consumption of buying a unit of the asset, while the right-hand side gives the expected future consumption benefit derived from the dividend and capital value of the asset. This relation plays a central role in modern theoretical models of optimal dynamic consumption and portfolio decisions, such as those of Robert Lucas.

*University of Pennsylvania and National Bureau of Economic Research, respectively. This research was supported by the National Science Foundation under grants SOC79-07651 and SOC79-13429. The views expressed here are not necessarily those of the supporting agency. Roger Huang provided research assistance.

Since $u'(C_t)$ and P_{it} are known at time t (in contrast to P_{it+1}, D_{it+1}, and C_{t+1} which are not), we can rewrite (1) as

$$1=E(R_{it}S_t|I_t) \tag{2}$$

where $S_t=\beta u'(C_{t+1})/u'(C_t)$ is the marginal rate of substitution between present and future consumption (the reciprocal of the usual measure), and $R_{it}=(P_{it+1}+D_{it+1})/P_{it}$ is the return (or rather one plus the rate of return as it is usually calculated). Note that the expectation in (2) conditional on information I_t is always 1. Hence it does not depend on I_t. Therefore, it equals the unconditional or simple expectation

$$1=E(R_{it}S_t) \tag{3}$$

Thus, the proper stochastic interpretation of the familiar two-period diagram is that the expected product of the uncertain return and the uncertain marginal rate of substitution is one. This means that $E(R_{it})$ needn't equal the subjective rate of time preference nor need it be the same for all assets ("expected profit opportunities" may exist). Instead, (3) says that a *weighted* expectation of returns, with weights corresponding to marginal rates of substitution, is the same for all assets. Returns which come in periods of low marginal utility of consumption (i.e., when consumption is high) are given little weight, because they do little good in terms of utility. Returns which come in periods of high marginal utility are given a lot of weight. The same expression can also be written another way, using the fact that the expected product of two variables is the product of their means plus their covariance:

$$E(R_{it})=E(S_t)^{-1}\cdot(1-cov(R_{it},S_t)) \tag{4}$$

Equation (4) states that the expected return of an asset depends on the covariance of the asset's return with the marginal rate of substitution. An asset is very "risky" if its payoff has a high negative covariance with S. (Douglas Breeden has recently persuasively argued for the use of consumption correlatedness as the appropriate measure of risk.)

The theory of asset returns embodied in each of expressions (1) through (4) is very powerful because it can be applied so generally. It holds for *any* asset, or portfolio of assets. It holds for *any* individual consumer who has the option of investing in stocks (even if he chooses not to hold stocks) and thus it must hold for aggregate consumption so long as some peoples' consumption is well represented by the aggregate consumption. It holds even if the individual's choices regarding other assets are constrained (for example, the individual cannot trade in his or her "human capital," is constrained by institutional factors in housing investment, or is unable to borrow money) so long as such constraints do not affect his ability to change his saving rate through stock purchases or sales. It incorporates all sorts of uncertainty that people consider in making investment decisions, since these factors are reflected in consumption. The model holds for any time interval, whether a month, a year, or a decade.

II. Perfect Foresight Stock Prices

By iterating (1), we find that price P_{it} at time $t<n$ is the expected present value of dividends and a terminal price P_{in} discounted by the marginal rates of substitution:

$$P_{it}=E\left[\sum_{j=1}^{n-t}\beta^j\frac{u'(C_{t+j})}{u'(C_t)}D_{it+j}+\beta^{n-t}\frac{u'(C_n)}{u'(C_t)}P_{in}|I_t\right] \tag{5}$$

It is useful to define the perfect foresight stock price P^*_{it}, which is the price at t given that the consumer knows the whole future time path of consumption, dividends, and the terminal price P_{in}:

$$P^*_{it}=\sum_{j=1}^{n-t}\beta^j\frac{u'(C_{t+j})}{u'(C_t)}D_{it+j}+\beta^{n-t}\frac{u'(C_n)}{u'(C_t)}P_{in} \tag{6}$$

Clearly (5) states that $P_{it} = E[P^*_{it} | I_t]$. Further, we assume that $u(C)$ is of the constant relative risk aversion form

$$(7) \qquad u(C) = \frac{1}{1-A} C^{1-A} \quad 0 < A < \infty$$

where A is the "coefficient of relative risk aversion," which is a measure of the concavity of the utility function or the disutility of consumption fluctuations.

Figure 1 shows a plot of P_t from 1889 to 1979, where P_t is the annual average Standard and Poor's Composite Stock Price Index divided by the consumption deflator. On the same figure, we plot the perfect foresight real price P^*_t for $A=0$ and $A=4$ using (6) and (7), where we use actual realized real annual dividends for the Standard and Poor series, the Kuznets-Kendrick-US *NIA* per capita real consumption on nondurables and services and the terminal date $n = 1979$. For each A, we generate a value of β so that (3) holds, as estimated by the sample mean. The case $A=0$ is revealing; this is the case of risk neutrality, and of a constant discount factor. Notice that with a constant discount factor, P^*_t just grows with the trend in dividends; it shows virtually none of the short-term variation of actual stock prices. The larger A is, the bigger the variations of P^*_t and $A=4$ was shown here because for this A, P, and P^* have movement of very similar magnitude. Irwin Friend and Marshall Blume estimated A to be about 2 under the assumption that the only stochastic component of wealth is stock returns. Irwin Friend and Joel Hasbrouck estimated A to be about 6 when stock returns and human capital are the stochastic components of wealth. We also computed a P^* series using after-tax returns. It did not look much different from the P^* shown here in the first half of the sample when income taxes were generally unimportant, and did not seem to fit P any better in the second half.

The rough correspondence between P^* and P (except for the recent data) shows that if we accept a coefficient of relative risk aversion of 4, we can to some extent reconcile the behavior of P with economic theory even under the assumption that future price movements are known with certainty. In a world of certainty, the marginal rate of substitution S_t would equal the inverse of one plus the real interest rate, ρ_t. Hence our equilibrium condition becomes $(P_{t+1} + D_{t+1}) \div P_t = 1 + \rho_t$. Thus it can be shown that real stock prices as well as real prices of other assets whose dividend is stable in real terms will rise dramatically over periods when real interest rates are very high. Real interest rates will be high when C_{t+1} is high relative to C_t, for example, in periods of depression when C_t is abnormally low. Hence it is an equilibrium for P_t to be low (relative to P_{t+1}) because otherwise people will desire to dissave (for example, by selling stock at t) in order to maintain their consumption level. Movements in real interest rates which are necessary to equilibrate desired savings to actual savings will lead to changes in stock prices even if dividends are unchanged. It is these movements which are brought out in the figure when P^* with $A=4$ is compared with P^* with $A=0$.

The correlation between P^* and P is perhaps not altogether surprising, given the correlation between the stock market and aggregate economic activity over the business cycle noted long ago by many people (see, for example, Arthur Burns and Wesley Mitchell). However, P^*_t is not merely a proxy for aggregate economic activity or consumption at time t. If we assume, as an approximation, that dividends follow a growth path $D_t = D_0 \delta^t$ and if we set $n = \infty$ in (6) to ignore the terminal price, then P^*_t is given by $P^*_t = D_0 \delta^t [C_t^A \Sigma_{k=0}^{\infty} (\beta\delta)^k C_{t+k}^{-A}]$. This says that P^*_t follows a growth path times the *ratio* of C_t^A to a weighted harmonic average of future C^A. The weights decline exponentially into the future. Thus, for example, P^* declines gradually between 1907 and 1919 not because consumption declined, since real per capita consumption remained more or less level over this period, but because the gap between current consumption and the longer-run outlook widened. In other words, P^* fell at this time because the perfect-foresight individual, knowing his economic fortune would eventually improve following

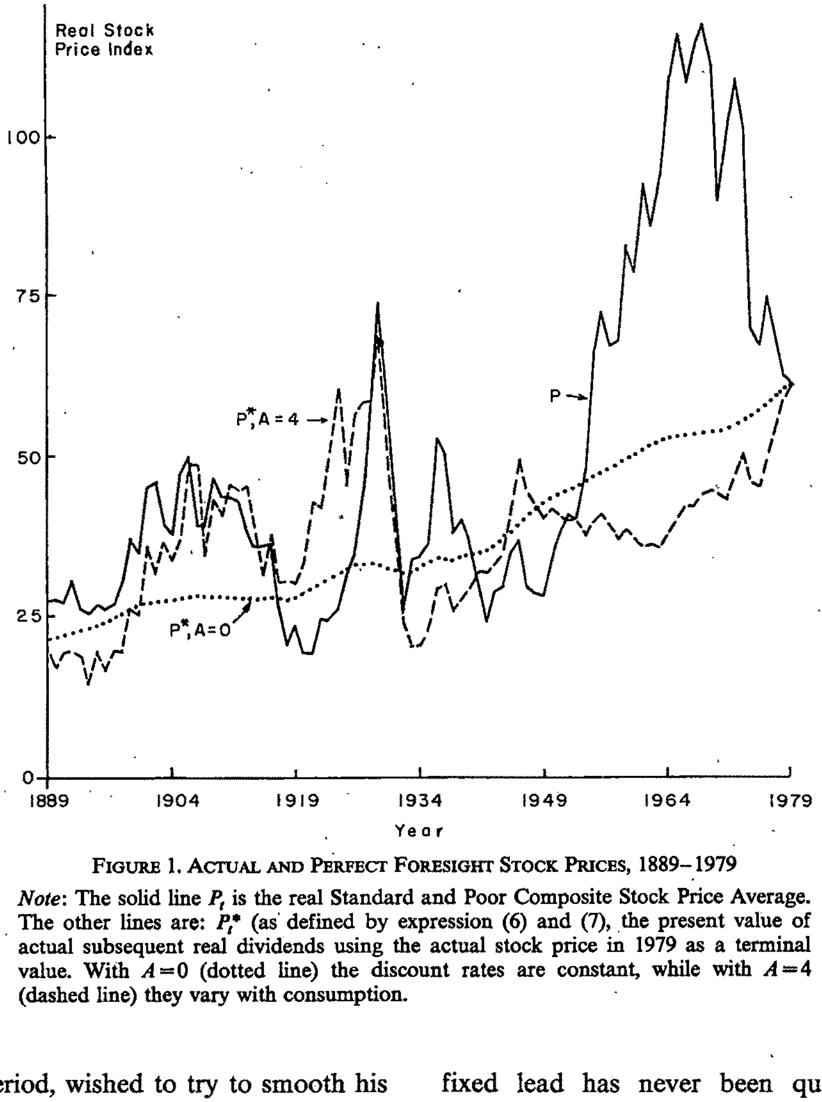


FIGURE 1. ACTUAL AND PERFECT FORESIGHT STOCK PRICES, 1889–1979

Note: The solid line P_t is the real Standard and Poor Composite Stock Price Average. The other lines are: P_t^* (as defined by expression (6) and (7), the present value of actual subsequent real dividends using the actual stock price in 1979 as a terminal value. With $A=0$ (dotted line) the discount rates are constant, while with $A=4$ (dashed line) they vary with consumption.

the war period, wished to try to smooth his consumption over this period. This kind of relationship between P and C would not have been visible by looking at raw stock price and economic activity index series alone, as the earlier scholars did. On the other hand, the short-run correspondence between P and P^* around such episodes as the panics of 1893 or 1907 was in effect noted by these authors.

Our construction implies that P^* (as well as P) is a leading indicator of future levels of economic activity, but it does not suggest the conventional notion of a fixed lead of a few months to a year between P and aggregate economic activity. However, such a fixed lead has never been quantitatively established (see C. W. J. Granger and M. Hatanaka).

Once we drop the assumption of perfect foresight, there need not be a close relationship between P_t and P_t^*. If consumers have no information about P_t^*, then P_t will be a constant and P_t^* will vary. We can write $P_t^* = P_t + U_t$ where $U_t = P_t^* - E[P_t^* | I_t]$ is a forecast error. Since P_t is in the information set I_t, U_t must be uncorrelated with P_t, so that the variability of the stochastic process $\{P_t^*\}$ will be *larger* than that of the stochastic process $\{P_t\}$. Further, if we consider any subset of the information set at t, say I_t^s, then $Var(P_t^* | I_t^s) \geqslant Var(P_t | I_t^s)$. If we make the assumption that the variability of the stochastic processes $\{P_t\}$ and $\{P_t^*\}$ can be estimated from the sample variability of observed P_t and P_t^*, then the figure can give some evidence in favor of the hypothesis that A is at least 4. From the figure, it is clear that with $A=0$ the variance inequality is reversed: P_t^* varies *less* than P_t. This is evidence against the hypothesis that the discount factory does not vary. Once we raise A to, say, $A=4$, then the variability of the discount factor forces P_t^* to vary a lot. The larger A is, the larger is the variability induced in P_t^* by changes in the consumption path. Another way that the reader can see that discount factor variability is important is to apply the above variance inequality with $I_t^s = D_t$, yielding $Var(P_t^* \mid D_t) \geqslant Var(P_t \mid D_t)$. If the discount factor was constant, then this states that current dividends should be a better predictor of the current stock price than current dividends can predict weighted future dividends. Casual observation suggests this is false. Current dividends are a very good forecaster of future dividends, and a terrible forecaster of the current stock price. Once we permit the discount factor to vary, the inequality has a much greater chance of being true, since the current dividend is a poor forecaster of future discount factors.

If it is accepted that the variability of the discount factor is important, then we can use this to provide evidence against Hall's assertion that short-term movements in consumption are not forecastable by consumers. To see this, write the jth term in the summation in (5) as $E(\beta^j u'(C_{t+j})/u'(C_t) \mid I_t)\, E(D_{t+j} \mid I_t) + cov(\beta^j u'(C_{t+j})/u'(C_t), D_{t+j} \mid I_t)$. If neither the expectation of $\beta^j u'(C_{t+j})/u'(C_t)$ nor its covariance with dividends is forecastable (depends on I_t), then this term varies only due to changes in the expectation of D_{t+j}, i.e., due to information about dividends. If, moreover, $E(\beta^j u'(C_{t+j})/u'(C_t) \mid I_t) = \gamma^j$ (as might be suggested by Hall's random walk hypothesis), then P_t equals $E(\hat{P}_t^* \mid I_t)$ where $\hat{P}_t^* = \Sigma\gamma^j D_{t+j}$ (plus a deterministic term due to the covariance). $\hat{P}_t^*$ has a *constant* discount factor and is proportional to P^* in Figure 1 with $A=0$. Because P_t^* with $A=0$ fails the variance test as mentioned previously, we tend to reject models with constant discount factors. Hence we conclude that consumption changes are forecastable. This implies that expected real interest rates vary (contrary to the claims of Eugene Fama and others).

This conclusion does not contradict Robert Hall's assertions that (i) to an *econometrician* who does not know as much as consumers, the marginal utility of consumption is a random walk, and (ii) that income may be a proxy for lagged consumption in econometric models which have shown that consumption is very sensitive to income. The fact that stock prices vary so much with consumption suggests that consumers have more information about consumption than is contained in current consumption, and this leads expected real interest rates to vary with information.

III. Further Research

We have some preliminary results on the estimation of A and β. Estimates of both parameters can be derived using expression (3) for two different assets, which we took as stocks and short-term bonds. Unfortunately, the estimates of A for the more recent subperiods seem implausibly high. This breakdown of the model mirrors the divergence between P^* and P since the early 1950's, as well as the extremely low real return on short-term bond rates in this period. There was an enormous rise in stock prices in that period which cannot be explained by changes in realized dividends or in marginal rates of substitution. Preliminary results show that it cannot be explained by taxes. Friend and Blume noticed an extremely high excess return of stocks over bonds in this period relative to all other subperiods from 1890 to date. Their estimated market price of risk was twice as high in the decade 1952–61 as the highest of any other decade. While the divergence between P_t and P_t^* might be considered an enormous forecast error, we don't have any idea as to why $E(P_t^* \mid I_t)$ should have changed so much.

REFERENCES

D. Breeden, "An Intertemporal Asset Pricing Model With Stochastic Consumption and Investment Opportunities," *J. Financ. Econ.*, Sept. 1979, *7*, 265–96.

Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles*, New York 1956.

E. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

I. Friend and M. Blume, "The Demand for Risky Assets," *Amer. Econ. Rev.*, Dec. 1975, *65*, 900–23.

______ **and J. Hasbrouck,** "Effect of Inflation on the Profitability and Valuation of U.S. Corporations," Univ. Pennsylvania 1980.

C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series*, Princeton 1964.

R. Hall, "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis," *J. Polit. Econ.*, Dec. 1978, *6*, 971–88.

S. LeRoy and R. Porter, "The Present Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, Mar. 1981.

______ **and C. J. La Civita,** "Risk Aversion and the Dispersion of Asset Prices," *J. Bus., Univ. Chicago*, 1981 forthcoming.

R. E. Lucas, "Asset Prices in an Exchange Economy," *Econometrica*, Nov. 1978, *46*, 1429–45.

R. Shiller, "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *Amer. Econ. Rev.*, June 1981, *71*.

MEASURING BENEFITS FROM AIR POLLUTION CONTROL: WHERE DO WE STAND?

What Do We Know about Benefits of Reduced Mortality from Air Pollution Control?

By SHELBY GERKING AND WILLIAM SCHULZE*

According to conventional wisdom, the main benefit of environmental regulation is improved health. Thus, research into the benefits of air pollution control has sought primarily to determine the extent to which morbidity and morality rates decline when air quality improves. Given a knowledge of this relationship, benefits of air pollution regulations can be estimated using the economic analysis of safety programs developed by such investigators as E. J. Mishan, Richard Thaler and Sherwin Rosen, V. Kerry Smith, and Brian Conley. The conceptual framework developed by these authors values small changes in risk using a willingness-to-pay measure, rather than the lost productivity (or earnings) from early death, and therefore avoids the numerous theoretical problems associated with the latter approach. However, the distinction between these two approaches to benefit estimation reaches far beyond purely theoretical considerations. For similar safety programs, estimates based upon willingness to pay measures are about ten times higher than those based upon productivity changes.[1]

Although progress has been made in valuing the benefits of improved health, the mortality effects of air pollution are less well understood, in spite of the claims of several statistical studies that a clear linkage exists. This paper argues that extraordinary difficulties are present in statistical epidemiology which have yet to be resolved. These difficulties arise in part because of problems in obtaining desirable data. Potential sources of information include first, controlled experimental data from either animal experiments or clinical trials, and second, uncontrolled data on human health and exposures in the real world. The first data source is of little use since the principle scientific questions are the health effects of long-term low-level exposures to air pollution which are impossible to simulate in a laboratory environment. Most biomedical authorities reject the notion that the health effects of high-level short-term exposures to air pollution can be extrapolated to low-level long-term exposures. An analogy can be made to use of table salt. Large sudden doses are deadly, while long-term low-level doses are necessary in the human diet.

This situation leaves the use of uncontrolled real world data on human health and exposures as the only game in town. Of course, economists have been quick to recognize the similarity of this epidemiological problem to many in economics which have been studied using statistical tools such as regression analysis. Use of ordinary least squares to attempt to account for uncontrolled factors and isolate the independent contribution of air pollution to human mortality has become quite popular (see work by Lave and Seskin, G. C. McDonald and R. C. Schwing, Allen Kneese and Schulze, Crocker, et. al.). However, with only a few exceptions, these studies have been unsophisticated in their application of econometric methods and have failed to look for, or cope with, a variety of potentially serious statistical problems.

*Associate professors, department of economics, University of Wyoming. We would like to thank the Environmental Protection Agency and, in particular, Alan Carlin for support for this research under contract R805059010. All remaining errors and conclusions are our sole responsibility.

[1] For example, Lester Lave and Eugene Seskin use about $30,000 as an average value of a life saved in increased productivity based on the work of Dorothy Rice in 1968. In contrast, T. Crocker et al. use $340,000 as the willingness to pay for an expected life saved based on the work of Thaler and Rosen.

The plan of the paper is to list a few of these problems in the next section and then to show how these problems can significantly affect estimated effects of air pollution on health using a data set consisting of mortality rates, air pollution levels and other variables for sixty *U.S.* cities.[2] Comments on policy implications are made in the conclusion.

I. Statistical Problems

The aim of this section is to outline some of the major statistical research problems that remain to be overcome in estimating the impact of air pollution on human health. These problems arise largely because the process by which air pollution affects health is not yet completely understood. As a result, any statistical specification of this relationship for the purpose of regression analysis is subject both to uncertainty and question. Most importantly, since the true model is not known with any degree of precision, the power of classical tests of hypotheses regarding the role of air pollution in causing illness or premature death is greatly diminished. To at least some extent, statisticians have faced difficulties of this general nature in virtually all areas of investigation. However, important environmental management decisions regarding air pollution control have been based, in part, upon regression equations where small changes in model specification appear to produce comparatively large changes in implications.

Because theoretical knowledge regarding the connection between air pollution and health is so inadequate, empirical efforts to identify this relationship must be interpreted with caution. Intuitively, there are at least three important types of specification error that should be thoroughly investigated prior to accepting present estimates for policy purposes: errors in functional form; omitted variables; and simultaneity. Clearly, these problems are not an exhaustive list of statistical difficulties in air pollution epidemiology research. Nevertheless, as will be argued momentarily, they do appear to lie at the root of many of the conflicting sets of estimates that have been obtained by other investigations. Each of these problems will now be considered in turn.

Economic and epidemiological theory provides few insights into the most appropriate functional form for a regression equation used to measure the impact of changes in air quality on human health. This situation is rather unfortunate since the true relationship between health and its determinants may be strongly non-linear. For example, the health consequences of changes in variables such as cigarette smoking, protein consumption, as well as air pollutants are likely to depend not only on the magnitude of the change, but also upon the levels of the variables themselves. Yet little is known about exactly how to specify these functional relationships. The issue of correct function form is important because benefit estimates are frequently obtained from simple equations where a mortality rate (or its natural logarithm) has been regressed on air pollution measures together with other explanatory variables (or their natural logarithms). In particular, these regressions are used to obtain the desired benefit estimates by making hypothetical changes in the air quality variables and then noting the effect on the health measure. Obviously, benefit estimates obtained by this procedure may be seriously biased unless these simple linear or *log*-linear functional specifications are accurate to a useful degree of approximation.

A second important consequence of the lack of information on the true air quality-health relationship involves the issue of omitted variables. As Henri Theil has shown, the error of mistakenly excluding variables from an otherwise correctly specified regression equation causes the estimated coefficients on all remaining included regressors to be biased and inconsistent. This issue is not unique to statistical work in the area under study; however, it seems particularly critical here because of apparent conflicts over the empirical determinants of mortality. On the one hand, previous investigations have shown significant adverse health ef-

[2] For a more complete examination of this data set, see Shultze, et al.

fects resulting from cigarette smoking and certain dietary habits. Nevertheless, when Smith analyzed thirty-two possible specifications of a regression equation (which are similar to those used by Lave and Seskin) where the dependent variable was the rate of mortality by *SMSA* and the explanatory variables were selected from among: 1) median age; 2) percent nonwhite; 3) population density; 4) temperature; and 5) particulates, little evidence of an omitted variables problem was found to be present. The *RESET* test, devised by James Ramsey, rejected the null hypothesis of a zero mean vector for the disturbance in only five of the thirty-two cases, while the *RASET* test failed to reject this null hypothesis in all cases. Because these tests were performed at the 10 percent level of significance and because their results may be unique to the particular data set employed, the appropriate role for other intuitively relevant variables in mortality rate estimating equations legitimately remains the subject of debate. Nevertheless, these results do lend support to the Lave and Seskin estimates of the impact of air pollution on health in the face of charges by other investigators including Crocker et. al. that they have omitted key mortality determinants.

Third, even though the results of Smith's *RASET* and *RESET* tests argue to the contrary, the estimation of an appropriately specified air pollution and health relationship may require the use of simultaneous equation estimation methods. Human decision making may cause the link between these two classes of variables to be considerably more complex than can be captured by a single equation. As an illustration, suppose that increases in medical care are effective in reducing mortality but that mortality rates exert an influence over where medical doctors and others in the health care field choose to locate. In this situation, a medical care variable should be included as an explanatory variable in a regression equation to explain the variation in mortality rates. Simple ordinary least squares estimation, however, may lead to biased and inconsistent estimates of all regression coefficients since the medical care variable would be correlated with the disturbance term even if the number of observations were arbitrarily large. A simultaneous equations estimation technique would be more appropriate in order to explicitly handle the problems created by this correlation.

In addition to the three factors just discussed, two less tractable, but no less important, research problems should be mentioned. First, as discussed by McDonald and Schwing, the variables used to measure air pollutants are often highly correlated with other explanatory variables. Because these pollutants are generated as joint products, in most cases, with other goods produced by the economic system, this situation should not be surprising. If the linear association between explanatory variables is high, separating the independent contribution of each to explaining the variation in mortality rates becomes difficult. McDonald and Schwing proposed a ridge regression estimator as a means of circumventing this problem. Ridge regression methods, however, are not entirely defensible as they represent a rather arbitrary, purely statistical solution to the multicollinearity problem and introduce a bias into the coefficient estimates that would not otherwise be present. (For a more complete critique of ridge regression procedures, see G. Smith and G. Campbell, together with various rejoinders to their paper.) Second, regression models are not highly sensitive and sophisticated research tools, particularly when the data used to estimate them contain measurement error. Such models may represent the best statistical tools available to social scientists. Nevertheless, they may not be up to the task of discerning the effect of air pollution on health when, in a correctly specified equation, other explanatory variables may be of much greater importance.

II. An Example

In this section, two tentative statistical models are presented in order to illustrate the importance of the problems relating to omitted variables and simultaneity that were raised in the previous section. Issues relating to such matters as the choice of functional

TABLE 1—DESCRIPTION OF DATA AND EMPIRICAL ESTIMATES

				Empirical Estimates[a]			
				MORT	*MORT*	*MDPC*	*MORT*
Variable	Year Units	Mean	*S.D.*	(1)	(2)	(3)	(4)
MORT	1970 Deaths/1000	11.283	2.161			5.823	
Total Mortality[b]						(1.392)	
MDPC Medical Doctors	1970 *MD*s/100,000	162.8	54.2		−.087		
per Capita[b]					(−5.764)		
NONW	1969 Fraction	.226	.154	2.997	9.996		2.349
Nonwhite Population				(2.403)	(6.389)		(2.365)
MAGE Median Age	1969 Years	28.82	2.74	5.73	.789		.626
of Population				(8.665)	(13.617)		(11.510)
DENS	1969% > 1.5	.022	0.013	12.940	49.794		18.217
Crowding in Homes	persons/room			(.881)	(3.934)		(1.447)
COLD	1972 no. days temp	86.9	47.7		0.21		.0175
Cold Weather	< 0°C				(4.468)		(3.421)
CIGS	1968 packs/yr/cap	165.8	23.25		.041		.00034
Cigarette Consumption					(4.693)		(.526)
PROT Animal Protein	1965 g/yr/cap	28,128.	1,603.4		.003		.00047
Consumption					(5.032)		(1.466)
CARB Carbohydrate	1965 g/yr/cap	123,490.	3,623.0		−.0001		−.00013
Consumption					(−2.366)		(−1.871)
SFAT Saturated	1965 g/yr/cap	16,315.	976.3		.0016		−.00068
Fatty Acids					(4.161)		(−2.616)
INCM	1969 $/yr/house-	10,763.	1,060.			.00925	−.000747
Median Income	hold					(1.143)	(−5.003)
EDUC	1969% > 25 yrs	55.3	7.4			.704	−.028
Education						(.616)	(−.893)
SO2X	1970 mg/m^3	26.92	22.2	.009	−.068	.070	.00118
Sulfur Dioxide				(1.059)	(−4.594)	(.192)	(.141)
PART	1970 mg/m^3	102.30	30.11	.011	−.015	−.514	.000184
Suspended Particulates				(2.006)	(−2.051)	(−2.085)	(.0374)
NO2X	1969 ppm	.076	.034	1.436	−11.081	−87.228	5.415
Nitrogen Dioxide				(.271)	(−2.332)	(−.381)	(1.238)
Constant				−7.719	−131.48	15.969	7.290
Degrees of Freedom				53.	47.	53.	46.
R^2				.692	—	—	.853
Estimation Method				OLS	2SLS	2SLS	OLS

[a] *t*-statistics are shown in parentheses.

[b] Predicted values, *MORT* or *MDPC*, are employed if these variables are used as explanatory variables in an estimated equation.

form and multicollinearity are not explicitly treated here, although they are certainly not less critical subjects for analysis. The first of these models, both of which are estimated using aggregate data on total mortality rates and other variables from sixty *U.S.* cities, is specified in the equation shown below.

$$(1)\quad MORT = F(NONW, MAGE, DENS, SO2X, PART, NO2X)$$

The exact definitions of all variables appearing in this equation, which are similar to those used by Smith, and Lave and Seskin, are presented in Table 1. In equation (1), variations in total mortality rates (*MORT*) are explained using variables measuring percent nonwhite (*NONW*), median age (*MAGE*), crowding (*DENS*), as well as the air pollutants (*SO2X*, *PART*, and *NO2X*). Ordinary least squares (*OLS*) estimates of this equation are presented in column (1) of Table 1 and *t*-statistics are presented beneath each coefficient estimate. These findings suggest that *SMSA*s with more older age residents, more nonwhites, and higher air pollution levels (especially in the form of particulates) have, in a statistical sense, significantly higher mortality rates at the 5 percent level. Examining only this equation, then, leads to the conclusion that air pollution kills people and that appropriate public policy measures should be taken to mitigate this hazard.

Rather different conclusions, however, are obtained from the statistical estimates of the second model. This model is specified in equations (2) and (3) and the exact definitions of all variables appearing there are given in Table 1.

$$(2)\quad MORT=g(MDPC, NONW, MAGE, DENS, COLD, CIGS, PROT, CARB, SFAT, SO2X, PART, NO2X)$$

$$(3)\quad MDPC=h(MORT, INCM, EDUC, SO2X, PART, NO2X)$$

Essentially, this structure builds upon equation (1). Equation (2) explains variations in *MORT* using variables including *NONW*, *MAGE*, and *DENS*, as well as *SO2X*, *PART*, and *NO2X*. But in addition, equation (2) also allows explicitly for the possibility that mortality rates are affected by cold temperatures (*COLD*) and by such lifestyle factors as cigarette smoking (*CIGS*), and diet (*PROT*, *CARB*, and *SFAT*), and by availability of medical care as measured by medical doctors (*MD*s) per capita (*MDPC*). Equation (3), then, hypothesizes that the location of *MD*s is determined by total mortality rates, *SMSA* income (*INCM*) and education (*EDUC*) levels as well as by the air quality variables.

Equations (2) and (3) are simultaneous in that variations in *MORT* are determined, in part, by variations in *MDPC* and vice versa. Due to this fact, and because under the order condition both equations appear to be identified, two-stage least squares (2*SLS*) is used as an estimation method. The estimates of these two structural equations are given in columns (2) and (3) of Table 1. With the exception of the coefficients on the air pollution variables, estimates of the slope parameters in equation (1) possess signs that might be expected on intuitive grounds. Increases in *MDPC* and in *CARB* contribute significantly to reductions in mortality rates, while colder *SMSA*s with more older-age residents, more nonwhites, more crowded housing conditions, and where more cigarettes are consumed tend to have higher mortality rates. These results suggest that *holding constant the linear influence of medical doctors per capita*, lifestyle variables measuring such factors as smoking and dietary habits exert a significant influence on total mortality rates; a finding that is of interest since variables of this type were ignored in specifying equation (1). On the other hand, the statistically significant but negative coefficients on the air pollution variables are rather more of a puzzle and cannot be completely explained. Nevertheless, a partial account of why this anomolous result has occurred will be offered momentarily. In the meantime, consider the estimates of the slope parameters of equation (3). According to these estimates, all but one of which are not statistically significant at conventional levels, medical doctors apparently avoid locating in *SMSA*s where particulate levels are high.

Additional insights into these results can be obtained by examining the estimates of the reduced form equation for *MORT*, which are shown in column (4) of Table 1. As indicated in the table, these estimates were obtained by applying ordinary least squares to an equation where *MORT* was specified to be a function of all exogenous variables in the structural model presented previously. There are two aspects of these estimates that are particularly worth noting. First, the estimates of the reduced form coefficients, unlike the structural coefficients, do not hold constant the linear influence of medical care and are interpreted as total, rather than partial, derivatives. In other words, the structural coefficients do not fully capture the fact that medical care may ameliorate the negative health effects of cigarette smoking, cold weather, crowded living conditions, and so forth. This ameliorative effect can only be determined by comparing the reduced form to the structural form coefficients. As is evident, such a comparison reveals that the coefficients on the socioeconomic and lifestyle variables are all smaller in the reduced form than in the structural form; a result suggesting that some ameliorative effects of medical care may indeed be present. Second, in the reduced-form mortality equation, the coefficients on the air pollution variables are positive. How can this result be explained? Although increased medical care would appear to reduce total mortality rates, doctors, according to the structural equation estimates, prefer not to live in polluted areas. Consequently, the reduced-form coefficients, which take this behavior into account, are large than their counterparts in the structural form. This observation, clearly, does not explain why the structural air pollution coefficients are negative. However, it does suggest that using reduced-form equations to measure the benefits of improvements in air quality may be somewhat misleading. In spite of the results from the structural model, reductions in air pollution may well reduce mortality rates. Nevertheless, as the reduced-form equation indicates, a portion of the reduction in mortality rates may result from improvements in medical care.

III. Conclusion

Existing statistical work on the mortality effects of air pollution has been interpreted to imply that control of stationary sources such as powerplants (which emit SO_2 and particulates) is justified while auto emission controls (particularly these for nitrogen oxides) are unjustified. These conclusions may be unwarranted for two reasons. First, as shown in the preceding section, the estimated effects of air pollution on human health are highly sensitive to model specification. With little or no a priori theoretical rationale for choosing one specification over another, a determination of the true health effects of air pollution is impossible. Future research, with primary data that is both collected specifically for the purpose of analyzing the health effects of air pollution and aimed at coping with the kinds of statistical problems identified here, may provide more convincing estimates. At the present time, however, relatively little is known about the effects of long-term low-level air pollution exposures on human mortality; certainly not enough to make benefit projections for policy purposes.

Second, the really important benefits from air pollution control may actually lie in the nonhealth area. For example, a recent study of the Los Angeles Basin suggested that a 30 percent reduction in ambient pollution levels (principally nitrogen oxides and related oxidant) would be worth nearly $1 billion per year to local residents (see D. Brookshire et al.). This study, using both a traditional hedonic property value study and survey questionnaires, concluded that a major fraction of perceived benefits were derived from the aesthetic (visibility and quality of life) benefits of reduced air pollution. Similarly, studies of the benefits of air pollution control in recreation areas such as the national parklands of the Southwest suggest that visibility and related nonhealth benefits are of principle concern. While supposed effects of air pollution on human mortality provide decision makers with an easy justification for control policies (often on ethical rather than economic grounds), economists ought to be concerned with all sources of benefits from pollution control on efficiency grounds. Serious doubt over the health effects of air pollution implies that less emphasis should be placed on health effects in making policy decisions.

REFERENCES

D. Brookshire, "Experiments in Valuing Public Goods," in V. Kerry Smith, ed., *Advances in Applied Micro-economics*, JAI Press, forthcoming.

B. Conley, "The Value of Human Life in the Demand for Human Safety," *Amer. Econ. Rev.*, Mar. 1976, *66*, 54–57.

T. Crocker et al., *Methods Development for Assessing Air Pollution Control Benefits*, Vol. 1, EPA-600/5-79-00/a, Feb. 1979.

A. V. Kneese and W. Schulze, "Environment, Health and Economics—The Case of Cancer," *Amer. Econ. Rev. Proc.*, Feb.

1977, *67*, 26–32.

L. B. Lave and E. P. Seskin, "An Analysis of the Association Between *U.S.* Mortality and Air Pollution," *J. Amer. Statist. Assoc.*, June 1973, *68*, 284–90.

Lester B. Lave and Eugene P. Seskin, *Air Pollution and Human Health*, Baltimore 1977.

G. C. McDonald and R. C. Schwing, "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, Aug. 1973, *15*, 463–81.

E. J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," *J. Polit. Econ.*, July/Aug. 1971, *79*, 687–705.

J. B. Ramsey, "Classical Model Selection Through Specification Error Tests," in Paul Zarembka, ed., *Frontiers in Econometrics*, New York 1974.

W. Schulze et al., "Mortality, Medicine, and Lifestyle," mimeo., Univ. Wyoming, Jan. 1980.

G. Smith and G. Campbell, "A Critique of Some Ridge Regression Methods," *J. Amer. Statist. Assoc.*, Mar. 1980, *75*, 74–81.

R. S. Smith, "The Feasibility of an 'Injury Tax Approach' to Occupational Safety," *Law, Contemporary Problems*, Summer-Autumn 1974, *38*, 730–44.

V. K. Smith, "Mortality—Air Pollution Relationships: A Comment," *J. Amer. Statist. Assoc.*, June 1975, *70*, 341–43.

R. Thaler and S. Rosen, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor E. Terleckyji, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1975.

H. Theil, "Specification Errors and the Estimation of Economic Relationships," *Rev. Int. Statist. Inst.*, No. 1-3, 1957, *15*, 41–51.

Measuring the Benefits from Reduced Morbidity

By M. L. CROPPER*

The predominant view in economics is that individuals are unaware of the health effects of air pollution and therefore do not take them into account in making decisions (see Lester Lave). Given this view, the appropriate way to measure the morbidity benefits of a reduction in pollution is to estimate a damage function and then assign a dollar value to the predicted decrease in illness. This, together with any reduction in medical costs, is what an individual would pay for a decrease in pollution if he treated his health as exogenous.

Unfortunately, this approach is inconsistent with the view, widely held in health economics, that individuals can affect the time they spend ill by investing in preventive health care. Support for this view is provided by Michael Grossman (1972a, b, 1976) whose work indicates that individuals diet, exercise, and purchase medical services to build up resistance to illness. These findings suggest that, if persons in polluted areas perceive their resistance to illness decreasing, they will try to compensate by exercising more, smoking less, or getting more sleep. Conversely, an improvement in air quality should lead to a decrease in preventive health care, and the value of this must be added to the benefits of pollution control.

Human capital theory thus implies that the damage function approach, by ignoring the value of preventive health care, understates willingness to pay for a change in air quality. This conclusion, it should be emphasized, does not assume that individuals know precisely the medical effects of air pollution. All that is necessary for a person to try and compensate for the effects of pollution is that he feels worse when pollution increases.

This paper presents a simple model of preventive health care, similar to that of Grossman (1972a, b), and uses the model to define what a person would pay for a change in air quality. The model assumes that one can build up resistance to acute illness by increasing his stock of health capital; however, health capital decays at a rate which depends on air pollution. For acute illness, willingness to pay as derived from the model is greater than the benefit estimate computed using the damage function approach. To illustrate the size of this discrepancy, estimates of willingness to pay are computed using data from the Michigan Panel Study of Income Dynamics.

I. A Model of Investment in Health

The essence of the human capital approach to health is that each individual is endowed with a stock of health capital H, with measures his resistance to illness. This stock can be increased by combining time TH_t with purchased goods M_t to produce investment in health,

$$(1) \qquad I_t = TH_t^{\,1-\zeta} M_t^{\,\zeta} E_{1t}^{\,\xi_1} \ldots E_{nt}^{\,\xi_n}$$

Outputs of equation (1) include exercise, rest, and nourishment. These will be affected by factors such as the individual's knowledge of health, or the presence of a chronic disease ($E_{1t}, \ldots, E_{nt}$ in equation (1)).

For simplicity, suppose that investment in health exhibits constant returns to scale so that the marginal cost of investment is constant and independent of I_t. This is reflected in equation (2) which gives the marginal cost of investment π_t as a function of the price of purchased goods PM_t, and the wage

*Assistant professor, University of Maryland and University of Southern California. This research was sponsored by NSF grant DAR8022374. I would like to thank Pauline Ippolito for comments on an earlier version of the paper, and Tom Crocker and Shaul Ben-David for providing the data. Amalia Arriaga-Salinas and Melanie Elliott provided excellent research assistance. Part of the computer time was provided by the University of Southern California and part by the Computer Science Center of the University of Maryland.

W_t,

$$(2) \qquad \pi_t = W_t^{\,1-\zeta} PM_t^{\,\zeta} E_{1t}^{\,-\xi_1} \ldots E_{nt}^{\,-\xi_n}$$

Investment in health increases the individual's health stock H_t, according to

$$(3) \qquad dH_t/dt = I_t - \delta_t H_t$$

Health capital also deteriorates at the proportional rate δ_t since resistance to illness would decline if no investments were made in health.

The main motive for investing in health is that health capital affects time spent ill, TL_t. For empirical work it is most appropriate to assume a threshold relationship between health capital and illness since a large number of persons (half of the Panel Study sample) report zero days of illness each year. A discontinuous relationship between H_t and TL_t, however, makes the solution to the individual's choice problem difficult. Let us therefore assume that the individual views the *log* of illness as a decreasing function of the *log* of health capital,

$$(4) \qquad ln\, TL_t = \gamma - \alpha\, ln\, H_t \quad \alpha > 0$$

This implies that time spent ill can be made arbitrarily small, although not zero.

Equations (3) and (4) suggest that the model, while appropriate for acute illness, should not be applied to chronic illness. In (4) a reduction in the health stock increases time spent ill; however, being ill in one instant does not reduce the stock of health capital in the next. This is reasonable only if TL_t refers to acute illnesses such as colds and the flu.

To simplify the model and facilitate estimation of willingness to pay (4) is assumed to be the only motive for investing in health. This reduces health to a pure investment good and implies that the only effect of health on utility is through the budget constraint.

In this case, the decision to invest in health can be separated from the decision to purchase other goods. First, a path of investment in health is chosen to maximize R, the present value of full income net of the cost of investment, then utility is maximized, given R. In the present model full income is the market value of the individual's healthy time. If Ω is the total time available at t, then $h_t = \Omega - TL_t$ is the amount of healthy time available. The present value of full income net of the cost of investing in health may therefore be written

$$(5) \qquad \int_0^T (W_t h_t - \pi_t I_t) e^{-rt}\, dt$$

where T is length of life. The individual's problem is to choose the path of investment which maximizes (5) subject to (3) and (4).

When the marginal cost of investment is constant, the solution to this problem is simple: at each instant the individual chooses an optimal level of resistance H_t^*, and then determines the amount to invest in health from (3).[1] The optimal health stock is determined by equating the value of the marginal product of health capital, $W_t \partial h_t / \partial H_t$, to its supply price,

$$(6) \qquad W_t \frac{\partial h_t}{\partial H_t} = \pi_t \left(r + \delta - \frac{d\pi_t}{dt} \frac{1}{\pi_t} \right)$$

The latter consists of three parts: the interest foregone by investing π_t in health rather than at the rate r; the depreciation cost $\pi_t \delta_t$, since each unit of health immediately declines by an amount δ_t; and a capital gain which accrues if the cost of investment is changing. If π_t is rising at approximately the rate of interest, then the right-hand side of (6) reduces to $\pi_t \delta_t$.

Substituting from (4) the optimal health stock may be written

$$(7) \qquad ln\, H_t^* = \frac{1}{1+\alpha} (\beta + ln\, W_t - ln\, \pi_t - ln\, \delta_t)$$

$$\beta = \gamma + ln\, \alpha$$

[1] For this solution to be valid, the resulting value of I_t must lie between 0 and $\bar{I}$, the maximum I permitted at any t. (That $\bar{I}$ exists is guaranteed by the fact that Ω and nonlabor income are finite.)

while time spent ill is given by

(8)
$$ln\,TL_t^* = \gamma - \frac{\alpha}{1+\alpha}(\beta + ln\,W_t - ln\,\pi_t - ln\,\delta_t)$$

There are several ways that pollution could enter this model. The observation that individuals are ill more often in polluted environments could mean that pollution enters the equation for time spent ill, (4), with a positive coefficient. This, however, implies that two individuals with the same health stock are not really equally healthy. Instead, it seems preferable to assume that pollution physically alters the state of a person's health.[2] This can be accomplished by making the rate of decay of health capital a function of air pollution P_t,

(9)
$$\delta_t = \delta_0 e^{\bar{\delta} t} P_t{}^{\psi} S_t{}^{\phi}$$

Equation (9) also implies that the rate of decay of health varies with age and with other factors, S_t, such as stress or pollution on the job.[3]

Adding equation (9) to the model means that it is more costly to build up resistance to illness in polluted environments, hence individuals in polluted areas will choose to maintain lower health stocks and will be ill more often than persons in cleaner areas. Proponents of the damage function approach might argue that this is unrealistic since individuals are unlikely to know the precise form of equation (9). All that is necessary, however, for an individual to choose a lower health stock is that he feels less healthy (perceives δ_t to be higher) when pollution increases. Knowing the precise relationship between δ_t and P_t is irrelevant in choosing H_t^*.

II. The Value of a Change in Air Pollution

Let us now consider the value to an individual of a small reduction in pollution at time t. Since a change in P_t affects net income only at t, the value of a small percentage change in P_t is defined as

(10)
$$-\frac{dR}{dP_t}P_t = \left(\frac{d\,ln\,TL_t}{d\,ln\,P_t}W_tTL_t + \frac{dI_t}{dP_t}\pi_tP_t\right)e^{-rt}$$

The first term on the right-hand side of (10) is the value of the reduction in sick time caused by a reduction in pollution. This is unambiguously positive. The second term describes the change in investment costs caused by a change in pollution. Reducing pollution increases the optimal health stock which, from (3), increases I_t^*. A reduction in P_t, however, also reduces δ_t which lowers the gross investment necessary to maintain a given health stock. For the functional forms above the net effect of these factors is positive, implying that a reduction in air pollution reduces resources devoted to preventive health care and thus increases willingness to pay,

(11)
$$-\frac{dR}{dP_t}P_t = \left(\frac{\alpha\psi}{1+\alpha}W_tTL_t + \frac{\alpha\psi}{1+\alpha}\pi_t\delta_tH_t^*\right)e^{-rt}$$
$$= 2\frac{\alpha\psi}{1+\alpha}W_tTL_te^{-rt}$$

If equation (10) is compared with the measure of benefits computed under the damage function approach, it is clear that the latter understates willingness to pay. Following Lave and Eugene Seskin, the damage function approach would measure the value of the reduction in sick time caused by a reduction in pollution, plus any change in medical costs. Since medical costs are negligible for acute illness, the damage function measure would equal the first term on the right-hand side of (10). The second term, which measures the decrease in resources devoted to preventive health care, would be

[2] It is also true that air pollution affects productivity of time spent exercising; however, not all time invested in health is affected in this way. It therefore seems inappropriate to incorporate pollution in the production function for health.

[3] In the paper δ_t is viewed as exogenous, hence the possibility of altering δ_t by moving or changing jobs is ignored.

ignored. To indicate the magnitude of this term and to give some idea of the morbidity costs of air pollution, I present estimates of (10) based on data from the Michigan Panel Study of Income Dynamics.

III. Estimation of Willingness to Pay

To compute willingness to pay requires an estimate of $\alpha\psi/(1+\alpha)$, the elasticity of sick time with respect to pollution. Equation (8) suggests that this can be obtained by regressing the *log* of sick time on the *log* of pollution and other variables which determine the optimal health stock. Since a large number of persons report zero days of illness each year, the appropriate statistical formulation of the equation is a Tobit model,

$$(12)\quad \begin{aligned} ln\,TL_{it} &= \text{undefined} && \text{if } X'_{it}B+u_{it}\leqslant 0 \\ ln\,TL_{it} &= X'_{it}B+u_{it} && \text{if } X'_{it}B+u_{it}>0 \end{aligned}$$

where

$$X_t=(1\; ln\,PM_t\, ln\,E_{1t}\ldots ln\,E_{nt}\, ln\,P_t\, ln\,S_t\, ln\,W_t\, t)$$

$$B'=\alpha(1+\alpha)^{-1}\big(\text{constant }1-\zeta-\xi_1\ldots$$

$$-\xi_n\psi\phi-(1-\zeta)\tilde{\delta}\big)$$

and $u_{it}\sim N(0,\sigma^2)$ for all t. Consistent estimates of (12) may be obtained by maximum likelihood.

Table 1 contains estimates of (12) for men between the ages of 18 and 45 from the Michigan Panel Study of Income Dynamics. The dependent variable is days lost from work due to illness, adjusted for differences in weeks worked. Independent variables, apart from the wage, either determine the rate of decay of health capital or affect the productivity of time invested in health.

Two features of the data should be noted. Since the dependent variable cannot be observed for persons too sick to work, the estimates in Table 1 are subject to selection bias. This problem is not serious, however, since only 3 percent of the sample is unable to work for health reasons. Secondly, the data support a threshold model such as (12) since approximately half of the sample reports zero days of illness each year.

Before computing willingness to pay, I comment briefly on the performance of the independent variables in Table 1. The first four variables measure factors which affect the rate of decay of health capital—air pollution, pollution at work, parents' income (which may affect δ_0), and race.[4] The first three of these consistently have the expected signs and are significant in six out of eight cases. Race, when significant, implies that being white increases the rate of decay of health capital. The second four variables affect the productivity of time spent investing in health. The presence of a chronic condition has a large negative impact on the productivity of time invested in health and is therefore positively related to sick time. Education, being married, and being cautious should increase the prevention received for a given expenditure of resources and are in most cases negatively related to illness.

The chief anomaly in the health equations is the behavior of the wage. A high wage, by increasing the value of healthy time, should increase H_t^* and reduce TL_t. In Table 1 the wage is either insignificant or positively related to illness. This could be caused by two factors. In the Panel Study the wage is computed by dividing labor income by hours worked. This is not a good measure of the marginal wage unless an individual receives the same wage for each hour worked. Secondly, as Grossman (1972b) has argued, the wage may act as a proxy for deleterious consumption habits, for example, eating rich food, which increase the rate of decay of health capital.

I turn now to estimates of willingness to pay. In Table 1 pollution is measured by the annual geometric mean of sulfur dioxide, which has been linked with acute illness in epidemiological studies. No other pollution variables are included since collinearity between pollutants leads to insignificant coefficients if several variables

[4] Age, which should also affect the rate of decay of health, was dropped from the equation for lack of significance.

TABLE 1—HEALTH EQUATIONS FOR MEN 18-45-YEARS OLD[a]

Independent Variable	Interview Year[b] 1970	1974	1976
Constant	3.5474	−1.2320	−0.5084
	(1.1253)	(0.9599)	(0.9014)
Ln(SO$_2$ Mean)	0.2879	0.3168	0.3189
	(0.2140)	(0.2076)	(0.1828)
Works in		0.5001	0.4828
Manufacturing[c]		(0.3659)	(0.3133)
Parents' Income	−0.1832	−0.1310	−0.0150
	(0.0936)	(0.1182)	(0.0953)
Race	0.7318	0.3768	−0.2950
(1=White)	(0.2697)	(0.4052)	(0.3084)
Has a Chronic	1.1972	0.6515	0.9347
Health Condition	(0.4582)	(0.2862)	(0.2602)
Years of Schooling	−0.1317	−0.1091	0.0496
	(0.0795)	(0.1170)	(0.0508)
Marital Status	−0.9678	0.9321	−0.6639
(1=Married)	(0.5098)	(0.4550)	(0.3828)
Risk Aversion	−0.3970		
Index[d]	(0.0881)		
Ln(Wage)	0.7492	−0.0899	0.1719
	(0.2873)	(0.3553)	(0.2813)
σ	2.1460	2.1586	2.1689
	(0.1824)	(0.2656)	(0.1931)
n	361.	247.	335.

Sources: All variables are from the Michigan Panel Study of Income Dynamics except SO$_2$ which is from the U.S. Environmental Protection Agency.

[a] The dependent variable in each equation is the *log* of [work-loss days/(days worked + work-loss days)] × 365. Standard errors appear beneath coefficients.

[b] Each interview year corresponds to the previous calendar year.

[c] Not available in 1970.

[d] Not available in 1974, 1976.

appear together. SO$_2$ should therefore be regarded as a pollution index and willingness to pay estimates viewed as indicators of the order of magnitude of willingness to pay. For the interview years 1970, 1974, and 1976, the mean of SO$_2$ is asymptotically significant at the .10 level or better (one-tailed test); furthermore its coefficient is approximately 0.3 in each year, despite differences in the specification of the health equation.

Consider now the amount an individual would pay for an x percent reduction in pollution. According to (11) this amount is

$$2(x/100)\frac{d\,ln\,TL_t}{d\,ln\,P_t}W_tTL_t \tag{13}$$

In equation (12) the elasticity of sick time with respect to pollution is equal to $\Phi(X_{it}'B/\sigma)$, the probability of being ill, times the coefficient of the *log* of pollution. Since $\Phi(X_{it}'B/\sigma)$ can be approximated by the fraction of the sample which is ill, $\Phi(X_{it}'B/\sigma)\doteq 0.5$ in each year, implying that the elasticity of sick time with respect to pollution $\doteq 0.15$.[5] The expected value of TL_t, calculated at the sample mean of X_{it}, is approximately 40 hours in each interview year.[6]

Equation (13) thus implies that the average person in the 1976 sample, who earned \$6.00 per hour, would pay \$7.20 annually for a 10 percent decrease in the mean of

[5] Evaluated at the sample mean of X_{it}, $\Phi(X_{it}'B/\sigma)=$ 0.57 in 1970, 0.50 in 1974, and 0.53 in 1976.

[6] $E(ln\,TL_{it})=X_{it}'B\Phi(X_{it}'B/\sigma)+\sigma\phi(X_{it}'B/\sigma)$. If this expression is evaluated at the sample mean of X_{it}, $E(TL_t)$ is, respectively, 46, 38, and 41 hours in 1970, 1974, and 1976.

SO_2. The damage function approach, by contrast, would put the value of a 10 percent reduction in pollution at only $3.60. In a city with one million prime-aged men, this would understate the value of a 10 percent reduction in air pollution by $3,600,000 annually. Ignoring adjustments to pollution, therefore, could sizably understate the value of an improvement in air quality.

REFERENCES

Michael Grossman, (1972a) "On the Concept of Health Capital and the Demand for Health," *J. Polit. Econ.*, Mar. 1972, *80*, 223–55.

______, (1972b) *The Demand for Health: A Theoretical and Empirical Investigation*, Nat. Bur. Econ. Res. Occas. Paper No. 119, New York 1972.

______, "The Correlation Between Health and Schooling," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1976.

Lester B. Lave, "Air Pollution Damage: Some Difficulties in Estimating the Value of Abatement," in Allen V. Kneese and Blair T. Bower, eds., *Environmental Quality Analysis*, Baltimore 1972.

______ **and Eugene P. Seskin,** *Air Pollution and Human Health*, Baltimore 1977.

U.S. Environmental Protection Agency, *Air Quality Data Annual Statistics*, Research Triangle Park, selected years.

University of Michigan Institute for Social Research, *A Panel Study of Income Dynamics, Procedures and Tape Codes*, Vols. 2, 4, 5, 6, Ann Arbor 1976.

Valuing Health Risk

By SHERWIN ROSEN*

In spite of the secular increase in longevity, the valuation of risks to human life has been an active area of economic research in recent years. There are three main elements of this general class of problems: the assessment and measurement of risk; the valuation of a given assessed risk; and behavioral responses to changes in inherent risk.

I. The Measurement of Risk

Throughout history most hazardous substances and activities were discovered by direct observation of use or action, with that knowledge filtering through the general population and ultimately getting reflected in actual practice. For example, the glazing material of pottery used for food no longer contains lead. While sharp observations linking cause and effect undoubtedly will continue to be the most prevalent form of detection, the use of chemicals today is so complicated that inferences about hazards are made with difficulty. Controlled experimentation and premarket testing are necessary supplements.

Analysis requires a conceptually systematic relation between exposure to risk and its probable effect on morbidity and mortality. The "dose-response" relationship suits this purpose. Dose is shorthand for exposure to risk, and response is the fraction of the exposed population suffering illness or death. This concept can be generalized to include all risks, both active and passive, say, bicycling and irradiation. It usefully illustrates that risk is not an all-or-nothing proposition, but rather can be controlled by varying exposure to risky situations. The fundamental economics problem is ascertaining the optimum degree of exposure by comparing costs and benefits. The benefit is expanded consumption opportunities, because incremental risk taking economizes resources that otherwise must be used to produce safety. The cost is the psychic disutility of additional risk bearing. The monetary equivalent of indirect utility costs is what is meant by the loaded term "value of a life."

Dose-response relationships are familiar to economists as statistical probit and logit functions. A simple but nonunique representation is to imagine the dose as an "insult" which depreciates one's health stock. Mortality occurs when the health stock falls below some critical level. The process is probabilistic because the initial health stock or the critical level (or both) are randomly distributed in the population at risk. As dose increases, depreciation increases, a greater proportion pass the critical level and the probability of an adverse outcome increases, sweeping out the underlying statistical distribution of latent stocks or thresholds.

Epidemiological studies establish dose-response relations for consumption and passive exposure risks in human populations, but are fraught with all the problems of nonexperimental data that economists can appreciate. Hence much effort has gone into estimating responses for nonhuman populations under controlled conditions. But there is no satisfactory substitute for human epidemiology because dose-response rates vary among species and cannot be extrapolated from one to another. For example, aflatoxin, a substance in a mold on peanuts and grain products, is one of the most potent carcinogens known. Yet dose-response rates at the median vary by three or four orders of magnitude among nonhuman species (see T. C. Campbell). Epidemiological studies in Africa and Asia, where peanuts are a cheap and important source of protein (not screened as thoroughly for aflatoxin as here), give a low-risk extrapolation to the

[*] University of Chicago and National Opinion Research Center. I am indebted to the National Science Foundation for financial support, and to Alan Garber and Richard Thaler for helpful discussions. Space limitations preclude references to all original sources.

U.S. population. Extrapolations from animal rather than human data shown human risks 18 times larger. Concluding that animal data systematically overstate risks to humans is unwarranted. There are as many examples where humans are sensitive and other species are not as vice versa, and it is impossible to predict the sign from a priori considerations. All that can be said is that a positive response in one species increases the likelihood of a positive response in others.

Much attention has been focused on risks at low doses and the existence of a threshold, defined as a dose below which risk is zero. The alternative hypothesis is linear response through the origin (no threshold). Search for a threshold is motivated by attempts to establish a "safe" exposure level below which further analysis is not required.

Different theoretical models of metabolic processes support one view or the other, but the issue is unresolved empirically. First, assessment of small risks requires immense amounts of (unavailable) data. Extrapolations outside the range of observations are sensitive to distributional assumptions and tail probabilities and standard errors of an extrapolation increase in distance from the sample mean. Second are the difficulties of extrapolating from one species to another. Third, there is little distinction between extensive and intensive exposure levels. The response differs depending on intensity of dose. Aflatoxin is a potent poison at high enough doses, but is carcinogenic when taken in low doses over an extended period of time. Penicillin causes death at high enough doses. Fourth, each person might have a threshold but it may differ among them. Given the present state of knowledge, I conclude that no absolute concepts of safety can be established and that cost benefit analysis cannot be avoided.

II. The Valuation of Risk

The natural concept in risk valuation is willingness to pay at the margin, the compensating variation. However, health risk has a characteristic which differentiates it from conventional goods because it is tied to consumption and work environments. Exposure can be varied by changing consumption levels, but no separate market for risk taking independent of actual consumption exists. One cannot divorce the utility of smoking from death risk. Hiring someone to smoke on one's behalf is useless. True, smoking less reduces risk, but it also reduces consumption value. Therefore the main analytical difference between valuing health risks and conventional goods is that no unique market price can be impersonally applied. Marginal valuations differ among persons and among activities.

Consider the simplest possible model in which the probability of surviving a single period is q. Utility conditional on survival is $U(C)$, where C is consumption. For a person with no heirs and therefore no demand for insurance, expected utility is $qU(C)$. The marginal rate of substitution between consumption and risk is the maximum amount a person will pay for a small increment of safety. It is $V \equiv -dC/dq = U(C)/qU'(C)$. Suppose in a group of N people risk is reduced by $1/N$. Each person will pay V/N for the reduction. The collective proceeds are $V = N \cdot (V/N)$ for an increase in expected survival of one person in the group: V is the value of life. It is readily verified that V decreases with q: those at greater risk have larger demand prices for safety. (This point appears often literature, but is most forcefully developed by William Gould and Richard Thaler and by Milton Weinstein et al.) Therefore maximizing the value of lives saved isn't necessarily the same as maximizing the number of lives saved. Lives at greater risk and with greater wealth weigh more heavily in this example. It follows that preference for ineffective, expensive crisis-oriented medical procedures rather than cheap cost-effective preventive measures affecting smaller risks can be rational. More is often invested in the health care of the sick (high risk) than in preventive measures for healthy people at small risk.

Given the interpersonal variation, it is not surprising that precise estimates of valuations are unavailable. A simplification is to express values in terms of human capital value lost from illness or death. This has the virtue of computation ease but is conceptually inappropriate. The relationship between human capital and V has been investigated by many people (see Joanne Linerooth). Expand the above example to include leisure, L. Then expected utility is $qU(C, L)$ to be maximized subject to $Y \equiv wT+I=C+wL$, where w is the wage rate, T is total time available, I is nonearned income and Y is full income. The solution conventionally equates the marginal rate of substitution with the wage rate. Substituting the demand equations into $U(\cdot)$ yields the indirect expected utility function on which a development similar to the one above gives $V=dY/dq=U/qU_c=(C/q)(U/U_cC)$. If nonlabor income is small enough to be ignored, then $C \approx wH$, where H is hours worked, and $V=(wH/q)(U/CU_c)$. Since q is approximately unity, willingness to pay exceeds total earnings if average utility exceeds marginal utility.

Strangely enough, whether V always exceeds market income depends upon whether the utility of death exceeds the utility of zero consumption! To illustrate, let $U=(C-\bar{C})^a(L-\bar{L})^b$ where $\bar{C}$, $\bar{L}$, a, and b are constants. Straightforward calculation gives $V=[(C-\bar{C})/Caq]wH$. When positive, $\bar{C}$ is conventionally described as the consumption level necessary for survival. Then willingness to pay exceeds earnings only when earnings are sufficiently high. For if $(C-\bar{C})/C$ is small enough, life isn't much worth living and earnings exceed V. If $\bar{C}$ is negative, life is worth living at all consumption levels, $(C-\bar{C})/Caq$ exceeds unity (risk aversion implies $a<1$) and the valuation of risk exceeds income. A similar development holds for $(L-\bar{L})/L$. The arcane and metaphysical nature of these concepts makes extrapolation to a threshold a trivial problem by comparison. Theoretical considerations alone therefore establish no operational connection between human capital and risk valuation. In fact, almost all empirical estimates of willingness to pay find that it exceeds income (see Glen Bloomquist).

Actual risk taking can be described as follows: Utility conditional on surviving is $U(C_1,\ldots,C_n)$, where $C_1,\ldots,C_n$ are n consumption goods. Consumption also affects survival probability: $q=q(a_1C_1,\ldots,a_nC_n)$ where $a_1,\ldots,a_n$ are nonnegative constants. For activities injurious to health, q_i (partial derivative) is negative. Others may increase longevity so that q_i is positive. A logit or probit model for q generates this specification. A linear index function in the C's interpreted as health stock leads to $q=q(a_1C_1+\ldots+a_nC_n)$ so a_i is related to the does-response rate for C_i. Alternatively, there may be several index functions, each corresponding to a separate cause of death, and q is not necessarily linear, as written above.

The budget constraint is $Y=\Sigma p_jC_j$, where p_j is the price of good j. Then necessary conditions for maximum expected utility are (compare A. Myrick Freeman): $a_iq_iU+qU_i-mp_i=0$, for each good, where m is the Lagrange multiplier associated with the constraint. It is plausible to assume that consumption of at least one good does not affect survival. Let it be C_n. Then $a_n=0$ and $m=qU_n/p_n$. Therefore $U_i/U_n+a_iq_i(U/qU_n)=p_i/p_n$. Recognizing from above that $V=U/qU_n$, then

$$U_i/U_n=p_i/p_n-a_iq_iV \tag{1}$$

Here a_iq_i is the marginal effect of consumption of good i on survival and V is the valuation of that increment. The rational consumer acts as a self regulator: $-a_iq_iV$ implicitly taxes goods that are injurious to health by raising their real prices above their money prices. It subsidizes goods that increase life expectency. Notice that the same valuation V is used by the consumer for all risks that are actually undertaken. The consumer uses the same marginal valuation for smoking, drinking, and living in a polluted neighborhood, though the marginal risk (a_iq_i) may be quite different among them.

This simple development clarifies the tie-in feature of risk and consumption and has implications for the actual estimation of V. The known price ratio in the marginal conditions is observed by the analyst and a_iq_i is known (in principle) from risk assessment studies. However, any imputation for V requires either isolating a class of instrumental goods for which direct marginal utility U_i is zero, or deciding a priori how preferences are split between direct utility and indirect longevity effects. Instrumental partitions of wants must be arbitrary at some point, and have no counterpart in modern economic theory. Value is whatever people are willing to pay; the reason and ultimate source of wants is immaterial. Due to these difficulties very few studies have inferred risk valuation from consumption patterns.

If a person won't consume risky goods which also yield disutility at positive prices, he may do so at negative prices. Herein is the intuition for imputations of neighborhood environmental disamenities from site values. Crime and pollution have been found to depress property values but no one has attempted to infer valuations of life from these estimates. Difficulties stand in the way. Estimates of mortality and morbidity risks associated with pollution across cities are suggestive, but are not sufficiently definitive to clearly assess health risks. Other, nonhealth aspects of pollution must be valued. This is doubly difficult empirically because uses of land and nonsite values interact with access and environmental attributes in important ways.

The main source of valuation estimates are wage differentials among risky jobs. Define a job risk index r and ignore leisure. Then $Y=Y(r)$ in the problem above, with $Y'>0$: riskier jobs must offer higher incomes or else they cannot be filled. Introducing r as an additional element of q, the maximum problem is augmented to include choice of r. The additional marginal condition is: $Y'(r)=q_r(U/qU_n)p_n$. In practice r is measured so that $q_r=-1$ so $Y'(r)/p_n=V$. The marginal risk premium is the value of life, so long as r is not an argument of $U(\cdot)$.

There are selectivity problems in actual data that lead to biased estimates. First, interpersonal variation leads people with smaller valuations to select the riskiest jobs. The estimate is biased downward for the average person (see my paper with Richard Thaler). This problem affects site value studies too. Second, if job risks are not statistically independent of consumption risks, or if there are unobserved differences among people that make real risks of any given work situation differ, another selection and corresponding bias is implied.

Actual estimates vary widely among studies, ranging from $250,000 in 1967 prices in studies using occupational wage differences to $1 million or more using industrial differences. Risk measurement error and different selectivity factors account for some of these differences. Occupational risks tend to be larger and therefore are likely to attract the less risk averse. They are also likely to be less error ridden. Yet no convincing reconciliation is available. This is a high priority problem on the research agenda. An interesting possibility is that each type of estimate contains differential components of consumption which are erroneously attributed to risk. For example, bartending and performing stunts are risky occupations but on-the-job consumption makes the measured money wage too small for the risk imputation in these cases. This is another example of difficulty in separating pure consumption and pure risk elements in (1).

The greatest barrier to a definitive study is major deficiency in the measurement of risks available for public use and public information. Incredibly, the main agency responsible for occupational safety and health requires firms to keep extensive records on injuries and accidents, but does not assess exposure to risk that would enable the calculation of probabilities.

III. Behavioral Responses to Changes in Safety

Changes in risk technology affect consumption patterns because opportunities are altered. This is incorporated in the model above by parametric changes in dose-response rates, a_i. Beneficial changes in a_i shift the survival distribution to the right for risky goods and to the left for healthful goods.

If C_i is risky, the marginal condition in (1) shows that a reduction in a_i reduces the implicit tax: $-a_iq_iV$ falls. Therefore the real price of C_i falls and the consumer substitutes in its favor. Exposure is increased to a smaller unit risk, tending to offset the parametric reduction. Detailed analysis shows that the actual amount of risk can increase after the change: orthopedic hospitals at the bottom of ski slopes may so encourage skiing that more fractures result. If pollution is abated, people spend more time outside. Improving crash and braking capacity of autos encourages greater speeds and more severe accidents, including possibly more pedestrian involvement (see Sam Peltzman).

The same manipulations as before on the indirect utility function give the marginal willingness to pay for small changes in a_i as $-dY/da_i = q_iC_iVp_n$. Willingness to pay is proportional to consumption. Though the greater exposure through changes in consumption tends to offset the first-order effects of changes in a_i, the value is always given by the expression for $-dY/da_i$ irrespective of how subsequent action affects q. Even if complete offset keeps q constant, consumers may be willing to pay substantial amounts for the improvement.

IV. Conclusion

The most pressing need is for better estimates of risk valuations. That will require much better data than are currently available. Also the role of the state in safety regulation has not been addressed here. No new issues arise for risks involving conventional externalities. Yet much safety legislation seems less concerned with externalities than with protecting us from ourselves. The case for government providing information about risks rests secure in the theory of public goods. Informational scale economies and problems of dissemination conceivably call for a principal-agent relationship in which certain types of decisions are delegated to the agency. However, the setting of standards and prohibitions is too simplistic. It seldom recognizes that personal action tends to interact with and partially offset them. Experience with drug lag and measurement of job-related risks give grounds for doubt about whether information is most efficiently provided in this way as well.

It is fashionable to claim conservatism by banning new substances and delaying others. But this may not be prudent once it is recognized that the status quo inevitably involves risks, benefits and costs as well. Much rhetoric would be avoided if it were clearly recognized that there are no absolute standards. It is always a comparison of one thing against another that is relevant. There is no free lunch in safety either.

REFERENCES

G. Blomquist, "The Value of Human Life: An Empirical Perspective," Illinois State Univ., Aug. 1979.

T. C. Campbell, "Aflatoxin Case Study," Report to the National Academy of Sciences, Institute of Medicine Food Safety Policy Study, 1978.

A. Myrick Freeman III, *The Benefits of Environmental Improvement,* Baltimore 1979.

W. Gould and R. Thaler, "Public Policy Toward Life Saving: Maximize Lives Saved Vs. Consumer Spending," Nat. Bur. Econ. Res., 1980.

J. Linnerooth, "The Value of Human Life: A Review of the Models," *Econ. Inquiry,* Jan. 1979, *17,* 52–74.

S. Peltzman, "The Effects of Automobile Safety Regulation," *J. Polit. Econ.,* Aug. 1975, *83,* 52–74.

R. Thaler and S. Rosen, "The Value of Saving a Life," in Nestor Terleckyj, ed., *Household Production and Consumption,* Nat. Bur. Econ. Res. *Stud. in Income and Wealth,* Vol. 40, New York 1975.

M. Weinstein, D. S. Shepard, and J. S. Pliskin, "The Economic Value of Changing Mortality Probabilities," *Quart. J. Econ.,* Mar. 1980, *94,* 373–95.

Federal Reserve System Implementation of Monetary Policy: Analytical Foundations of the New Approach

By STEPHEN H. AXILROD AND DAVID E. LINDSEY*

On October 6, 1979, the Federal Reserve announced a change in its open-market operating procedures that moved the focus of short-run guides for open-market operations away from the federal funds rate and toward reserve aggregates. Under the old procedures, the Trading Desk maintained close week-to-week control over the federal funds rate within a range specified by the Federal Open Market Committee (*FOMC*). Alteration of the funds rate between *FOMC* meetings was subject to guidelines conditional on behavior of the money supply relative to specified tolerance ranges. Under the new procedures the Trading Desk targets on a family of reserve aggregates—principally nonborrowed and total reserves. The reserve operating guides under the new procedures are derived from targets for growth of the monetary aggregates and are expressed as averages of weekly levels over the short-run operating period, normally the intermeeting period. As a corollary, less attention is paid to the federal funds rate, and it is permitted to vary over a much wider range during the short-run operating period as compared with the old procedures.

It should be understood that the October shift to a reserve operating guide for open-market operations did not also encompass a change in either the proximate or ultimate objectives of monetary policy; they continued to be, respectively, the money supply and goals for economic growth, employment, and prices. Rather, the shift involved a change in the technique of achieving the proximate money supply objectives. Two reasons can be advanced for the shift. First, the old technique of operating on the federal funds rate seemed to be a less reliable means, in practice, of attaining the proximate policy objectives. Second, in view of the worsening inflationary psychology at the time, a shift to a different technique of operations would help to dramatize the Federal Reserve's continuing commitment to slowing money growth to curb inflation, and thereby would work to reduce inflationary psychology.

This paper first summarizes the analytical framework underlying the choice between a reserve aggregate and a federal funds rate operating target, as well as some previous research bearing on the issue. Next, the technical aspects of implementing the new procedures are explained and contrasted to the theoretical paradigm underlying previous empirical results. Finally, experience with the new procedures since last October is reviewed and some conclusions are drawn.

I. Reserves vs. Interest Rates for Controlling Money

Theoretically, money supply can be controlled with a federal funds rate operating guide or with a reserves guide. Under a reserve operating target, and abstracting from the complications introduced by lagged reserve accounting in the very short run, the money stock is determined by the interaction of money supply and demand functions, with a short-term interest rate, such as

*Staff Director for Monetary and Financial Policy, and Assistant Director, Division of Research and Statistics, respectively, Board of Governors of the Federal Reserve System. The views expressed herein do not necessarily represent the views of the Board of Governors of the Federal Reserve System or other members of its staff.

the federal funds rate, serving as the endogenous price variable. The control variable is often taken to be nonborrowed reserves, though the role of monetary base has also been the subject of many studies. The following theoretical discussion is in terms of a nonborrowed reserve operating target, though much of the analysis also applies to a total reserves or monetary base target (with certain modifications, such as those needed for member bank borrowing).

Random disturbances can displace the money supply function from its expected position for any given interest rate. The error term for this function represents unexpected variations in banks' demands for excess and borrowed reserves; in demands by banks and the public for reservable bank liabilities not included in the money stock such as interbank, government, and large time deposits; and in the composition of assets in the money supply as between those with relatively high or those with relatively low reserve requirements. Similarly, the demand function can be displaced from its expected position for any given interest rate by an error term that incorporates the uncertainty built into the function itself and unforeseen deviations of real income or prices from expected levels. The level of the nonborrowed reserves operating target is set in advance to produce the targeted level of the money stock when the supply and demand functions are in their expected positions. Given this level of nonborrowed reserves, the impact on the money stock of either a supply-side or a demand-side error will be muted by partially offsetting movements in interest rates. The degree of offset to each type of shock depends on the interest elasticities of the two functions.

Under a federal funds rate operating target, the money stock is demand determined. Nonborrowed reserves become the endogenous variable that reacts to disturbances in order to maintain the predetermined federal funds rate. Given the inverse partial relationship between the federal funds rate and the quantity of money demanded, a higher federal funds rate implies a lower money stock, other things equal. The level of the federal funds rate operating target is set to produce the targeted money stock, given specific forecasts of real income and prices and, thus, the expected position of the demand function. The effect of a supply side disturbance on the money stock is fully offset by an induced change in nonborrowed reserves accomplished through open-market operations. In that case, the funds rate target contributes to the attainment of the money supply objective. However, if money demand varies from expectations, adherence to a funds rate target would mean that the disturbance would be fully accommodated by a change in nonborrowed reserves that would support a growth in money that deviates from the objective.

With theoretical analysis ambiguous as to the best control instrument, empirical evidence is needed to help determine the choice between the two methods. Such evidence comes from analyses by a number of economists that are designed to determine whether better predictions of money can be obtained from a reduced form equation that involves reserves and incorporates both the supply and demand functions or from the money demand function.[1] For example, published work by James Pierce and Thomas Thomson, Richard Davis, and Charles Sivesind and Kevin Hurley, as well as unpublished work by the Board staff, draws upon this abstract framework as a basis for examining the potential closeness of short-run monetary control available under the

[1] While theoretical considerations alone are unable to determine whether a funds rate or a reserves operating target could provide for closer short-run monetary control, it turns out, somewhat counterintuitively, that assuming known parameters and contemporaneous reserve accounting, a reserves target is unambiguously superior to an interest rate target in the instructive special case in which the error terms of the supply and demand functions have equal variances and a zero covariance, regardless of their relative interest elasticities. With a zero convariance, the error variance of the reduced form is a weighted sum of the error variances of the supply and demand functions; the interest elasticities determine the weights, which sum to less than one. If, in addition to these conditions, the interest elasticities also are equal in absolute value, a pure reserve target coincides with the optimal combination policy formulated by William Poole. See Stephen LeRoy and Lindsey for a derivation of these results.

two procedures. These studies use regressions of monthly data to compare the standard errors of estimated money demand functions with those of estimated reduced forms for money.

These studies thus focus primarily on control over the very short run of a month. However, in practice money objectives cannot be expected to be closely attained over such an horizon. Given the looseness of the control mechanism under the present institutional structure, or even under any feasible structure, it would be virtually impossible to attain precise control over money week-to-week or month-to-month. Moreover, in light of irreducible uncertainties about the appropriate definition and seasonal adjustment of money, together with the large random component in the public's short-run money demand, such precise short-run control over a particular measure of the money stock would not necessarily be desirable even if it were possible. Instead, the money supply objectives are best viewed over an horizon of three months or more. Even with that horizon, variations in growth rates from quarter-to-quarter (sometimes substantial variations) can be expected in response to factors noted above and other circumstances. In this context, the fundamental problem of monetary control reduces to one of ensuring that shorter-run deviations of actual from targeted monetary growth are offset over time by sufficiently slow, or rapid growth later, as the case requires.

The focus on control of money growth over the longer run does not mean, however, that the characteristics of operating procedures over the shorter run are irrelevant. Indeed, it might be argued that the desired long-run outcome is most likely to be attained by *aiming* in the short run at growth rates close to the long-run desired outcome, adjusted to take account of previous deviations of actual from desired growth or perhaps of known special circumstances (such as a large tax rebate by the government). Thus, the empirical evidence about whether money can best be controlled over a short-run operating period of a month or so with a nonborrowed reserves or a federal funds rate operating target becomes relevant.

In that context, the results of the studies examined suggest a virtual draw. For example, in the most recently published study, covering the years 1969 to mid-1974, Sivesind and Hurley find that the standard error of the M-1 demand equation was 3.2 percentage points, expressed as an annualized monthly growth rate, compared to a 3.3 percentage-point standard error for its reduced form. Other work cited above, which also includes evidence about the base and total reserves, reaches similar conclusions.

We believe, however, that, without careful interpretation, these results can give a misleading impression of the relative merits in practice of the two procedures for attaining money supply objectives, to the detriment of the reserves-oriented technique. In particular, inferences about monetary control from these results assume that the instrument is fully readjusted each period to a setting consistent with the interim target for the money stock. For the federal funds rate target, this would in practice mean substantial month-to-month changes. However, over the years in which the Federal Reserve System was targeting on the funds rate, it was not common for the rate to exhibit large, discrete changes immediately or soon after *FOMC* meetings. Instead, the federal funds rate, although evidencing substantial fluctuations over the business cycle, tended to move fairly smoothly from month to month. The apparent unwillingness of the *FOMC* to permit sharp changes in the funds rate over very short periods under the old procedures probably contributed to larger divergences of actual from desired money growth than would be suggested by the standard errors of money demand functions discussed above. This unwillingness may have reflected in part the inherent caution of policymakers in adjusting the monetary instrument (whatever instrument they choose) in light of uncertainties about economic processes.

When reserves are the policy instrument, however, a prompt response in the federal funds rate to changes in money demand is automatically permitted—without the need for a collective decision by the central policymaking body—that acts to forestall prolonged, cumulative departures of money growth from target. Indeed, the Federal Reserve noted in its announcement of the new procedures that the funds rate would be expected to vary more freely in response to market forces. The sharp increase in rates during the late winter of this year accompanying rapid money growth, followed by interest rate declines of unprecedented speed when the money aggregates weakened in the spring, is evidence of this effect at work.

Another practical relative advantage of the reserves operating technique, as compared with empirical results, is the ability in practice to establish the reserves target on the basis of recent evidence about shifts in the multiplier between reserves and deposits. Presumably, knowledge of recent public preferences for deposits, by type, and bank behavior with respect to demands for excess reserves and for borrowing from the discount window, should permit more accurate specification of a reserves target.

II. Implementation of New Procedures

Implementation of the new procedures has been designed to bring out the practical advantages of a reserve technique as compared with a funds rate procedure. In addition to the relatively wide funds rate range that has been established, the new reserve targeting procedure involves judgmental estimates, updated monthly and weekly, of the relationship between reserves and money, rather than reliance on multipliers implicit in regressions that are subject to drift over time. The reserve targets that guide open-market operations after an *FOMC* meeting are derived from the shorter-run objectives for the monetary aggregates set at that meeting. These objectives usually cover a period of several months, and are established as growth rates deemed to be consistent for that period with the longer-run, one-year growth rates that are the announced proximate objectives of policy.

Once the *FOMC* sets the shorter-run objectives for various monetary targets—such as *M*-1A, *M*-1B, and *M*-2—averages of weekly not seasonally adjusted levels of nonborrowed reserves, total reserves, and the monetary base are then established for the period between *FOMC* meetings consistent with these money stock targets, given estimates of the individual components of the money stock, of other deposits that may absorb reserves (such as interbank deposits), and of banks' demand for excess reserves and borrowing. These levels can be based on a constant seasonally adjusted rate of growth of the money targets on, say, a month-by-month basis or can involve variable monthly growth rates within the target period if there are clear reasons for it (for instance, it may be known that first weeks of a period are already relatively low so that the first month is more reasonably targeted to be relatively low with later months commensurately higher). (See Board of Governors of the Federal Reserve System.)

While the monetary aggregates are in a mechanical sense more closely related to total than to nonborrowed reserves (since it is total reserves that support deposits), only nonborrowed reserves, however, are directly controllable in the short run through open-market operations, although with some error arising from uncontrolled factors, such as float. Thus, while total reserves represent the principal reserve objective over an entire intermeeting control period, on a week-to-week basis nonborrowed reserves serve as the primary operating target.[2]

The present system of lagged reserve accounting reinforces the weekly focus on nonborrowed reserves, because required reserves are predetermined in any week by the level of reservable deposits two weeks earlier. Since excess reserves desired by banks are normally minimal, banks' demands for total reserves are virtually fixed in a given week, abstracting from reserve

[2] The monetary base is given less weight in operations than total reserves in order to minimize disturbances arising from unanticipated movements in currency, its principal component. If currency is running stronger than expected, achievement of a predetermined base target would require a dollar-for-dollar weakening in member bank reserves, causing a multiple contraction of bank deposits and money. In contrast, achievement of a predetermined total reserves target would imply that the money stock would be stronger than expected, but only by the amount by which currency is stronger than anticipated. Of course, once the unanticipated behavior of currency was recognized, compensating adjustments could be made to either a total reserves or monetary base operating target.

carryover. As a result, an attempt by the Trading Desk to reduce total reserves below this level demanded by banks through operations that lower nonborrowed reserves would be frustrated by an offsetting increase in member banks' borrowing at the discount window. Given increasing bank reluctance to make added use of the window as borrowings expand, the funds rate would have to rise sufficiently to induce just enough borrowings to fill the gap between total reserves demanded and the supply of nonborrowed reserves. Similarly, an attempt by the Trading Desk to increase total reserves in a given week above the level demanded by banks by increasing nonborrowed reserves would be frustrated by exactly offsetting reductions in bank borrowings. At some point borrowings would fall to zero, nonborrowed reserves would begin to exceed demanded total reserves, the funds rate would be driven sharply downward, probably violating the *FOMC's* lower bound, and open-market sales of securities to withdraw reserves may be necessary. Even with contemporaneous accounting, total reserves would be more difficult to control than nonborrowed reserves, since banks could, to an imperfectly predictable extent, alter discount window borrowings as a substitute for the portfolio adjustments necessary to move systemwide required reserves, and hence reserves demanded, by the desired amount. Thus, nonborrowed reserves are used, and necessarily used under the existing institutional structure, as the primary operating target in the short run.

Modifications are made to the nonborrowed reserve target, though, to compensate for certain disturbances. On the supply side, if early in the control period it appears from incoming data that the average level of interbank deposits, for example, will diverge from projected levels for the entire intermeeting period, then a compensating adjustment is made to the nonborrowed and total reserves paths. Similarly, adjustments can be made for unexpected variations in excess reserves or, in the case of the nonborrowed path, in borrowings when it is clear that there has been a shift in the demand for borrowings, given the discount rate, market rates, and the money supply. Such adjustments are made cautiously, however, to avoid overreaction to transitory variations.

In addition, adjustments also can be made to the targeted nonborrowed reserve path in response to unexpected changes in the quantity of money demanded by the public. For example, if the automatic pressures resulting from an unwanted strengthening of deposits and required reserves do not appear early in the control period to be adequate to induce the monetary aggregates and total reserves to move back toward the targeted paths as rapidly as desired, then the nonborrowed reserves target can be lowered to speed up the process. In this case, the tightening of money market conditions that automatically occurs when required reserves increase and banks bid more intensively for federal funds would be amplified by the reduced supply of nonborrowed reserves; as a result, the public would more promptly be induced to adjust the quantity of money demanded, and banks would more promptly adjust portfolio and lending policies, both reactions working to reverse the initial strengthening in deposits.

While such adjustments tend to improve monetary control, limitations on the federal funds rate could require a departure from reserve paths. Whether they do in practice, of course, depends on the willingness of the *FOMC* to adjust the funds rate range when the upper or lower limits are reached. The substantial variation in the funds rate since October reflects the willingness to establish an initially wide range of tolerance, and also a tendency to change the limits when they are approached and appear inconsistent with reserve and monetary targets.

III. Evaluation and Concluding Comment

Experience with the new procedures appears relatively satisfactory so far. Money growth has generally been within desired longer-run ranges, despite substantial month-to-month gyrations. Interest rates have been responsive to changes in market forces and have moved more promptly in a contracyclical direction, though not without fits and starts and a certain amount of market confusion and uncertainty, partly reflecting the natural process of adaptation to a new environment. In addition, inflationary expectations have at least been kept from worsening, and may be abating, despite such developments as slippage in budgetary restraint.

Since October 1979, money growth has shown considerable variation over the short run, but has tended to return to within longer-run ranges after moving outside them. During 1980, the levels of *M*-1A and *M*-1B rose near to and a bit above the upper end of their respective long-run ranges in February, but by March were near the midpoints of the ranges. *M*-2 was within its range through March. Then record monthly declines in *M*-1A and *M*-1B in April pushed all the monetary aggregates well below the lower bounds of their ranges. Subsequently, *M*-2 returned to its lower bound in May, and *M*-1B in July. Published weekly data for August suggest that *M*-1A will be above its lower bound, *M*-1B above the midpoint of its longer-run range, and *M*-2 close to the upper end of its longer-run range. Thus, the extraordinary weakness of the aggregates during April was fairly promptly reversed. At the same time, as noted earlier, market interest rates, after having risen substantially in late winter, moved sharply lower, probably much faster than would have occurred under a funds rate procedure, as the economy weakened. Moreover, during the second quarter, it appeared that the narrow monetary aggregates may have been experiencing a renewal of the downward demand drift that occurred in the mid-1970's—perhaps in response to the very high and record levels of short-term interest rates reached in early spring—justifying some shortfall at the time relative to the midpoints of the ranges.

While money has been reasonably well controlled on average since the new procedures have gone into effect, the large variations in money growth over short-run periods suggest three possibly conflicting conclusions. First, they might indicate that the control mechanism needs improving. Second, they might suggest that there is enough noise in short-run money demand so that efforts to control money closely in the short run may be either unavailing or undesirable. And third, the sharp drop of money growth in the second quarter raises the question of the stability of money demand over a longer run and therefore the question of the appropriateness of predetermined monetary objectives.

Many economists could easily point to institutional reforms that would in their view improve the precision of monetary control through reserve targeting. Some reforms are in process. Institutional arrangements are soon to be altered dramatically by the implementation of the Monetary Control Act of 1980. By increasing the coverage of Federal Reserve requirements to all depository institutions, by making required reserve ratios more uniform on transactions balances, and by eliminating reserve requirements on all nontransactions balances except nonpersonal time deposits, the Act, once fully phased in, should improve the precision of control over the narrow monetary aggregates through reserve targeting, despite the expected increase in holdings of "excess" reserves in the form of vault cash at small banks. As an additional step in this direction, the Board is studying the practicability of a return to contemporaneous reserve accounting in 1981.

Other reforms that could be mentioned include a penalty discount rate, at the extreme closing the discount window (except for emergency loans), reverse lagged reserve accounting, staggered reserve settlement systems, equal reserve requirements on all deposits included in the measure of money to be controlled, and to reach far back in years 100 percent reserve requirements on demand deposits and other assets that are deemed to be equivalent to currency for carrying out transactions. It would take another paper to discuss all of these and others that might be advanced, together with their advantages and disadvantages.

We might on that subject simply offer our summary view that, given recent experience, we see little need for substantial changes in the monetary system for purposes of monetary control other than those in process. Partly we take this view because the large amount of short-run money demand noise in the financial system makes the need for precise short-run money supply control technically questionable. The development of money market funds is worrisome, however, and whether at least those funds used actively for transaction purposes should not be subject to Federal Reserve requirements is a real question. It is appealing to have equal reserve requirements on all assets included in the measure of money to be controlled—if agreement could be reached on such a measure—since that would clearly simplify the multiplier problem. But we would argue that the reserve requirement level should be sufficiently high so that it was generally "binding" on financial institutions, that is, at that level which led to required reserves at least as high as the great bulk of banks would in any event maintain for operating purposes.[3]

In general, we believe these institutional questions that affect the precision of the reserves-to-money relationship are minor compared with the basic issue of whether reserve targeting linked to money supply objectives is the best method for carrying out monetary policy. Experience to date suggests an affirmative answer for the economic conditions since October 1979. But a reserve targeting procedure linked to predetermined money growth rates assumes a more or less stable demand function for money. However, as more and more substitutes for money evolve, as different forms of money develop, and as financial technology becomes more and more computerized and transfers for payments out of almost any and all assets can be made rapidly by electronic means, it may become increasingly difficult to detect—indeed, to believe in—a stable demand function for money. This raises the question of whether we will not eventually reach a point where interest rates will have to be given more consideration in policy, in the absence of a clear notion about what is "really" money and in view of the possibility that the velocity of whatever we happen to define as money may come to develop the capacity for varying sharply from period to period.

[3] Ideally, all such required reserves should receive a market-related interest return in order to forestall innovations designed to avoid reserve requirements, but this step does not appear to be politically feasible.

REFERENCES

S. H. Axilrod, "Monetary Aggregates and Money Market Conditions in Open Market Policy," *Fed. Reserve Bull.*, Feb. 1971, 79–104.

______ **and D. Beck,** "Role of Projections and Data Evaluation with Monetary Aggregates as Policy Targets," in *Controlling Monetary Aggregates II: The Implementation*, Fed. Reserve Bank Boston Conference Series, No. 9, Sept. 1972, 81–102.

R. G. Davis, "Implementing Open Market Policy with Monetary Aggregates Objectives," *Monetary Aggregates and Monetary Policy*, Fed. Reserve Bank New York, 1974, 1–19.

S. F. LeRoy, "Monetary Control Under Lagged Reserve Accounting," *Southern Econ. J.*, Oct. 1979, *46*, 460–70.

______ **and D. E. Lindsey,** "Determining the Monetary Instrument: A Diagrammatic Exposition," *Amer. Econ. Rev.*, Dec. 1978, *68*, 929–34.

J. L. Pierce and T. D. Thomson, "Some Issues in Controlling the Stock of Money," in *Controlling Monetary Aggregates II: The Implementation*, Fed. Reserve Bank Boston Conference Series, No. 9, Sept. 1972, 115–36.

W. Poole, "Optimal Choice of Monetary Policy in a Simple Stochastic Macro Model," *Quart. J. Econ.*, May 1970, *84*, 197–216.

C. Sivesind and K. Hurley, "Choosing an Operating Target for Monetary Policy," *Quart. J. Econ.*, Feb. 1980, *94*, 199–203.

Board of Governors of the Federal Reserve System, "Appendix B: Description of the New Procedures for Controlling Money," appended to "Monetary Policy Report to Congress Pursuant to the Full Employment and Balanced Growth Act of 1978," Feb. 19, 1980.

Monetary and Fiscal Policies in an Open Economy

By JACOB A. FRENKEL AND MICHAEL L. MUSSA*

The central theme of this paper is that international linkages between national economies influence, in fundamentally important ways, the effectiveness and proper conduct of national macro-economic policies. Specifically, our purpose is to summarize the implications for the conduct of macro-economic policies in open economies of both the traditional approach to open economy macroeconomics (as developed largely by James Meade, Robert Mundell, and J. Marcus Fleming) and of more recent developments. Our discussion is organized around three key linkages between national economies: through commodity trade; through capital mobility; and through exchange of national monies. These linkages have important implications concerning the effects of macro-economic policies in open economies that differ from the effects of such policies in closed economies.

Recent developments in the theory of macro-economic policy have established conditions for the effectiveness of policies in influencing output and employment which emphasize the distinction between anticipated and unanticipated policy actions, the importance of incomplete information, and the consequences of contracts that fix nominal wages and prices over finite intervals. In this paper, we shall not analyze how these conditions are modified in an open economy. However, since our concern is with macro-economic policy, a principal objective of which is to influence output and employment, we shall assume that requisite conditions for such influence are satisfied.

I. Commodity Market Linkages

International trade links the prices of goods produced and consumed in different national economies. This linkage has at least three implications for the conduct of macroeconomic policy in open economies.

First, according to the principle of purchasing power parity, the price level in one country (in terms of domestic money) should equal the price level in a foreign country (in terms of foreign money) multiplied by the exchange rate between domestic money and foreign money. Because of transport costs, trade barriers, different weighting schemes for price indices and changes in relative prices of nontraded goods, this link is not rigid; but the evidence indicates that this principle holds fairly well over long time periods (though it has weakened during the 1970's). The key implication of purchasing power parity is that a country cannot choose its long-run inflation rate independently of its long-run monetary policy and the long-run behavior of its exchange rate. A country, particularly a small country, that fixes the exchange rate between its domestic money and the money of some foreign country will experience a domestic inflation rate and a domestic rate of monetary expansion that are strongly influenced by the monetary policy of that foreign country. This is so even if changes in real economic conditions (which are largely independent of domestic monetary policy) induce divergences from strict purchasing power parity.

Second, the world monetary system and the conduct of national monetary policies must allow for changes in equilibrium relationships between national price levels induced by changes in relative prices of internationally traded goods and of nontradable goods. To maintain a system of fixed exchange rates, changes in equilibrium relationships among national price levels must

*University of Chicago and National Bureau of Economic Research, Inc., and the University of Chicago, respectively. Frenkel acknowledges research support by the National Science Foundation, grant SOC 78-14480. This research is part of the National Bureau of Economic Research Program in International Studies. The views expressed are our own and not necessarily those of the NBER.

be accomodated by differentials among national inflation rates, supported by appropriate national economic policies. Under a system of controlled or managed floating, it is essential that countries either allow their inflation rates or the rates of change of exchange rates to accommodate equilibrium changes in relative national price levels. A rule that links changes in the exchange rates rigidly to changes in domestic and foreign prices, in accord with relative purchasing power parity, is not consistent with this requirement.

Third, macro-economic policy can do little to offset changes in equilibrium levels of real income resulting from changes in relative prices of internationally traded goods. A case in point is the recent increase in the relative price of oil. Monetary policy can influence the extent to which the increase in the *relative* price of oil affects general price levels and perhaps short-run levels of employment in oil exporting and oil importing countries. Tax and expenditure policy can affect the extent to which gains and losses of real income are translated into changes in real expenditure, or are financed by changes in foreign lending and borrowing. By influencing the level and distribution of real expenditure, fiscal policy can also affect the relative prices of nontradable commodities and the distribution of the change in real national income among individuals within the economy. However, neither monetary nor fiscal policy can alter to any appreciable extent the average change in the long-run level of real expenditure resulting from a change in the relative prices of internationally traded commodities that are beyond the control of national economic policies.

II. Capital Market Linkages

International capital mobility links interest rates on financial assets denominated in different national monies through the interest parity relationship. This relationship requires that interest differentials between securities denominated in different currencies equal the forward discount or premium on foreign exchange. The empirical evidence indicates that this relationship holds almost exactly for easily tradable securities identical in all respects except currency of denomination, but somewhat less well for assets exchanged primarily in national credit markets (see Robert Aliber, Michael Dooley and Peter Isard, and Frenkel and Richard Levich). International capital mobility also allows countries to finance imbalances in their current accounts and thus provides an important channel for the international transmission of macro-economic disturbances. The linkage of interest rates through interest parity and the transmission of macro-economic disturbances through international capital flows have significant implications for the conduct of macro-economic policy in open economies.

International capital mobility imposes a severe constraint on the use of monetary policy for domestic stabilization purposes. Under a fixed exchange rate, an increase in the domestic credit component of the money supply in a small open economy may temporarily reduce interest rates on domestic securities, but will induce a capital outflow and a corresponding loss of foreign exchange reserves that will rapidly reduce the money supply back to its previous equilibrium level. Monetary expansion by a large country, which affects conditions in world financial markets, can be somewhat more effective in influencing domestic prices, output, and employment. However, even a large country will suffer a loss of foreign exchange reserves that is inversely related to its size in the world economy. Sterilization policies of a central bank may temporarily insulate the domestic money supply from changes in foreign exchange reserves; but, in the long run, sterilization cannot sustain a money supply that differs from the equilibrium level of money demand. Under a flexible exchange rate, a government regains long-run control over the nominal money supply. However, international capital mobility still limits the effectiveness of monetary policy: Any increase in aggregate demand induced by lower domestic interest rates is partially dissipated in increased expenditures on imported goods, financed by international capital flows; and exchange rate adjustments that occur rapidly in response to perceived changes in monetary policy are likely to lead to rapid adjustments of domestic prices and wage rates, thereby limiting the effect of monetary policy on output and employment.

A high degree of capital mobility also implies a low degree of effectiveness of fiscal policy. Under a flexible exchange rate, a fixed domestic money supply and a domestic interest rate fixed by conditions in world markets (and by exchange rate expectations which affect the forward discount or premium on foreign exchange) impose a strict constraint on the level of domestic income that is consistent with monetary equilibrium. Fiscal policy actions do not affect this constraint (except possibly by altering exchange rate expectations) and, hence, cannot affect the equilibrium level of domestic income. Under a fixed exchange rate, the money supply is not fixed because the capital inflow induced by an expansionary fiscal policy will increase the foreign exchange reserves of the monetary authority. The initial expansionary effect of any fiscal stimulus, however, is limited by the extent to which it falls on domestically produced goods that are not close substitutes for imports; and the subsequent multiplier effects of any fiscal stimulus are limited by the high marginal propensity to spend on internationally traded goods.

To achieve the maximum effect from fiscal and monetary policy in open economies, it follows that such policies should be directed toward goods and assets that are isolated from world trade, that is, toward goods and assets for which the home country is "large" relative to the size of the market. Changes in government expenditures on nontradable goods are likely to be more effective in influencing domestic output and employment than changes in government expenditure on internationally traded goods. Similarly, open-market operations involving financial assets that are not close substitutes for international financial assets are more likely to influence interest rates and thus other macro-economic variables (see Rudiger Dornbusch, William Branson, and Russell Boyer). This does not imply, however, that it is desirable to artificially restrict trade and capital movements in order to enhance the effectiveness of macro-economic policy. Such restrictions have an important cost in terms of reducing the benefits that a country derives from integration of markets. Moreover, substitution possibilities among goods and among financial assets limit the effectiveness of restrictions on trade and capital movements. As many policymakers have discovered, using macro-economic policy for domestic stabilization objectives in an open economy is like trying to heat a house when the doors and windows are open and the cold wind is blowing.

Finally, international capital mobility implies that current account imbalances can be financed by capital movements, independent of the government's willingness to allow changes in its foreign exchange reserves. Hence, with capital mobility, the current account which measures both the net contribution of the foreign sector to aggregate demand for domestically produced goods and the change in a country's net debtor position, is always a concern of macroeconomic policy, regardless of the exchange rate regime. As emphasized by the absorption approach to the balance of payments (see Sidney Alexander and Harry Johnson), the current account is equal to the excess of national income over national expenditure. It follows that the policies required to bring about an improvement in the current account are not primarily commercial policies that affect the domestic relative prices of imported goods, but monetary and especially fiscal policies that stimulate private saving and that reduce the government's own excess of spending over income.

III. Monetary Linkages

The concept of monetary equilibrium as requiring equality between the demand and supply of money, is a basic ingredient in both closed and open economy macroeconomic theory. It implies that any change in the supply of money or any exogenous disturbance to money demand must lead to changes in the equilibirum values of one or more of the variables that influence money demand. It also implies that any disturbance or policy action that does not directly affect the demand or supply of money must, in equilibrium, lead to offsetting changes in the variables that influence money demand that are consistent with a constant level of that demand.

The implications of monetary equilibrium for the macro-economic policy of an open economy operating under a fixed exchange rate are reflected in the principles of the monetary approach to the balance of payments (see Mussa, 1974, and the articles in Frenkel and Johnson, 1976). Specifically, the use of monetary policy for domestic stabilization purposes is constrained by the equilibrium level of the demand for money which is largely beyond the control of the monetary authority. An expansion for the domestic credit component of the money supply may temporarily raise prices (especially of nontraded goods), raise output (especially in industries with sticky wages and prices), and reduce interest rates (especially for domestic securities sheltered from world financial markets). In the longer run, however, the direct effect of monetary expansion on desired spending and on desired portfolio reallocations, and the indirect effects of changes in prices and interest rates will induce deficits in the current and capital accounts of the balance of payments. Hence the foreign exchange reserve component of the money supply will decline until the money supply is reduced to the long-run equilibrium level of money demand. For a large country, the long-run equilibrium level of money demand may be influenced by the effect of domestic monetary expansion on the world price level. However, except for a reserve currency country, whose national money is accepted as a foreign exchange reserve by other countries, the principal long-run effect of monetary policy will be on the composition of assets on the central bank's balance sheet rather than on the magnitude of its monetary liabilities. Further, other economic policies affect a country's foreign exchange reserves (and hence its cumulative official settlements balance) only to the extent that they affect the demand to hold domestic money. For example, an import tariff can induce a once-and-for-all increase in the level of reserves to the extent that it increases the domestic price level and thus the demand to hold domestic money. But, it cannot induce a continuing inflow of foreign exchange reserves by reducing imports relative to exports (see Mussa, 1974).

The requirements of monetary equilibrium also constrain economic policy under a flexible exchange rate. A flexible exchange rate is not an additional policy tool that can be manipulated by the government, but rather an endogenous variable that is determined by market forces which are influenced by the actual and expected conduct of fiscal and, especially, monetary policy. In particular, from the homogeneity postulate, it follows that, other things constant, in the long run, changes in the nominal money supply will lead to proportionate changes in all nominal prices, including the price of foreign exchange. This is one of the fundamental tenets of the monetary approach to exchange rates (see Frenkel and Johnson, 1978).

Other important implications of this approach follow from the essential dynamic linkage between current exchange rates and expectations of future exchange rates implied by international mobility of financial assets denominated in different national monies.[1] First, since future government policies will influence future exchange rates, it follows that expectations concerning future policies should influence current exchange rates. Hence, the effect of any particular policy action on exchange rates (and through exchange rates on other macroeconomic variables) will depend on its effect on expectations concerning future policy actions. Second, the sensitivity of exchange rates to expectations of future policy implies that the traditional approach to macroeconomic policy analysis, which views policy

[1] The emphasis on new information which alters expectations concerning future exchange rates and thereby induces an immediate adjustment of current exchange rates is fundamental in explaining the recent volatility of exchange rates; see Mussa (1976, 1979), Dornbusch (1978), Frenkel (1981), and Frenkel and Mussa.

as isolated actions in response to particular circumstances, is inappropriate. Instead, it is necessary to analyze the general framework of government policy, within which the effect of any particular action depends on the public's perception of the implications of that action for the future conduct of policy. Third, since exchange rates respond quickly to new information about events likely to affect foreign exchange markets, exchange rate adjustments are an important channel for rapid transmission of macro-economic disturbances and of government policies. In particular, new information that leads to the expectation of a higher rate of inflation is likely to induce an immediate depreciation of the foreign exchange value of domestic money which will be rapidly translated into increased domestic prices of internationally traded commodities. In addition, exchange rate depreciation may serve as a signal for upward adjustments in the prices of domestic goods and in wage rates. Fourth, if there is short-run stickiness of prices of domestic goods in terms of national monies, then rapid exchange-rate adjustments will induce changes in the relative prices of different national outputs. Such relative price changes are a desirable response to changing real economic conditions requiring adjustments of relative prices; and a flexible exchange-rate regime may have an important advantage in facilitating such adjustments. However, unnecessary changes in relative prices in response to purely monetary disturbances may have significant social costs that can be avoided or reduced by pursuing monetary policies that are not themselves an independent source of monetary disturbances and that offset, as much as possible, exogenous fluctuations in money demands. Fifth, official intervention in foreign exchange markets, which alters only the supplies of nonmonetary assets available to the public, may have a limited influence on exchange rates through portfolio balance effects. In addition, such intervention may have a more powerful effect on exchange rates by signalling to the public the intentions of governments concerning future policies. Finally, exchange rates may be useful as an indicator for monetary policy directed at offsetting fluctuations in money demand, especially when rapidly changing inflationary expectations make nominal interest rates an unreliable indicator of fluctuations in money demand. For example, the combination of rising nominal interest rates and appreciation of the foreign exchange value of domestic money would probably indicate an increase in the demand for domestic money that should be accommodated by an increase in supply; whereas the combination of rising interest rates and depreciation would probably indicate an acceleration of inflationary expectations that should not be fueled by an accommodative monetary policy.

IV. Concluding Remarks

In this paper we have discussed the implications of "openness" of an economy for the effectiveness and appropriate conduct of macro-economic policies. As recent experience demonstrates, no country is immune from disturbances originating in the rest of the world, and no government can sensibly conduct its macro-economic policy on the assumption that it operates in a closed economy. National economies are linked not only through the mechanism of the Keynesian foreign-trade multiplier but also through the complex of linkages implied by commodity trade, capital mobility and the exchange of national monies. These linkages are not properties of a particular mode but implications of parity conditions and equilibrium requirements in goods and asset markets, income and balance sheets constraints, the absence of long-run money illusion, and consistency of expectations which impose important constraints on the conduct of macro-economic policy in an open economy.

REFERENCES

S. S. Alexander, "Effects of a Devaluation on a Trade Balance," *Int. Monet. Fund Staff Papers*, Apr. 1952, *2*, 263–78.

R. Z. Aliber, "The Interest Rate Parity Theorem: A Reinterpretation," *J. Polit. Econ.*, Nov./Dec. 1973, *81*, 1451–59.

R. S. Boyer, "Commodity Markets and Bond Markets in a Small, Fixed-Exchange-Rate Economy," *Can. J. Econ.*, Feb. 1975, *8*, 1–23.

W. H. Branson, "Portfolio Equilibrium and Monetary Policy with Foreign and Non-Traded Assets," in Emil Claassen and Pascal Salin, eds., *Recent Issues in International Monetary Economics*, Amsterdam 1976.

M. P. Dooley and P. Isard, "Capital Controls, Political Risk, and Deviations from Interest-Rate Parity," *J. Polit. Econ.*, Apr. 1980, *88*, 370–84.

Rudiger Dornbusch, "Capital Mobility and Portfolio Balance," in Robert Z. Aliber, ed., *The Political Economy of Monetary Reform*, Montclair 1977.

______, "Monetary Policy Under Exchange Rate Flexibility," in *Managed Exchange Rate Flexibility: The Recent Experience*, Fed. Reserve Bank Boston Conference Series, No. 20, 1978.

J. M. Fleming, "Domestic Financial Policies Under Fixed and Under Floating Exchange Rates," *Int. Monet. Fund Staff Papers*, Nov. 1962, *9*, 369–79.

Jacob A. Frenkel, "Flexible Exchange Rates, Prices and the Role of 'News': Lessons from the 1970's," *J. Polit. Econ.*, Aug. 1981, *89*.

______ **and Harry G. Johnson,** *The Monetary Approach to the Balance of Payments*, London; Toronto 1976.

______ **and** ______, *The Economics of Exchange Rates: Selected Studies*, Reading 1978.

______ **and R. M. Levich,** "Transactions Costs and Interest Arbitrage: Tranquil versus Turbulent Periods," *J. Polit. Econ.*, Dec. 1977, *85*, 1209–26.

______ **and M. L. Mussa,** "The Efficiency of Foreign Exchange Markets and Measures of Turbulence," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 374–81.

Harry G. Johnson, "Towards a General Theory of the Balance of Payments," in his *International Trade and Economic Growth*, Cambridge 1958.

James E. Meade, *The Theory of International Economic Policy, Vol. I: The Balance of Payments*, London 1951.

Robert A. Mundell, *International Economics*, New York 1968.

M. L. Mussa, "A Monetary Approach to Balance-of-Payments Analysis," *J. Money, Credit, Banking*, Aug. 1974, *6*, 333–51.

______, "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy Under a Regime of Controlled Floating," *Scand. J. Econ.*, May 1976, *78*, 229–48.

______, "Empirical Regularities in the Behavior of Exchange Rates and Theories of the Foreign Exchange Market," in Karl Brunner and Allan Meltzer, eds., *Policies for Employment, Prices, and Exchange Rates*, Vol. 11, Carnegie-Rochester Conferences on Public Policy, *J. Monet. Econ.*, Suppl. 1979, 9–57.

Rational Expectations and the Conduct of Monetary Policy

By ANDREW F. BRIMMER AND ALLEN SINAI*

A controversial challenge to the traditional formulation of economic policy and use of large-scale macro-econometric models in policy analysis is "rational expectations" (*RE*). The elements of *RE* that bear on monetary policy include:

1) *Natural rate hypothesis*: (*NRH*) suggests that the relationship between real magnitudes and their natural levels depends only on unanticipated inflation. Thus, unless changes in monetary policy cause actual inflation to deviate from expected inflation, no effect on the real economy can occur.

2) *Efficient markets*: The rational expectations hypothesis (*REH*) assumes that information gathering by the private sector is efficient, that is, the public is aware of and uses all relevant information in forming expectations. If so, and given the correct structural model of the interaction between monetary policy and the economy, there can be no surprises from the implementation of monetary policy. The true parameters underlying economic processes are conditioned by efficient markets.

3) *Ineffectiveness of stabilization policy*: Under *RE*, systematic and known economic policy cannot affect real output, employment, or inflation. Economic agents understand the implications of policy and anticipate them because information is processed efficiently in all markets. With no surprises possible, systematic monetary policy is ineffective.

4) *Irrelevance of large-scale macro-econometric models for the analysis of monetary policy*: Since macro-econometric models contain only estimates of true historical structural parameters, effects of new changes in policy are not captured. Therefore, large-scale macro-econometric models cannot effectively analyze the impacts of economic policy. In effect, equations are missing from econometric models that connect varying structural parameters to changes in policy.

5) *Knowledge of feedback rules*: Corollary to *RE* is that knowledge exists on feedback from the economy to policy. This suggests that monetary policy will be ineffective unless the public can be kept ignorant of the feedback rules.

This paper examines several questions relative to *RE* and the conduct of monetary policy: 1) what do the notions of *RE* imply for monetary policy? 2) Was the radical policy shift adopted by the Federal Reserve on October 6, 1979 and the resulting impacts consistent with *RE*? 3) Are arguments against the use of large-scale macro-econometric models in policy analysis valid or could the results from model simulations be robust with respect to the *RE* criticism?

In particular, the New Fed Policy (*NFP*) begun in October 1979 is analyzed as *if* it had been motivated by *RE*. There is much to suggest such an interpretation. The *NFP* was completely unanticipated, the new policy rules have not been readily comprehended, and elements of surprise still persist in practice. Thus, the adoption of the *NFP* represents the equivalent of a laboratory experiment where some of the key elements in *RE* can be tested.

Such an examination is conducted through simulation of the *NFP* in the Data Resources, Inc. (DRI) Quarterly Model of the U.S. Economy. Of course, the coefficients in the DRI model do not yet reflect enough experience under the *NFP* to assess its long-run effects. However, history does reflect the impacts of Federal Reserve policy on interest rates and of interest rates on portfolio adjustments, the real economy, and inflation. And, to the extent that expectations are incorporated into the structure of

*Brimmer and Company, Inc. and Data Resources, Inc., respectively.

the DRI model, it should be less inefficient as an instrument for policy analysis.

Thus, we also survey the role of expectations in the DRI model, both extrapolative and rational. Although pure *RE* do not yet appear, elements are modeled for some markets. And, from the model simulations of the *NFP*, we conclude that large-scale macro-econometric models might be robust with respect to policy analysis, despite possible shortcomings in the modeling of expectations. This conclusion arises from uncertainty on the applicability of *RE* to the real world and the significant impact on the economy found for the *NFP* in the DRI model simulations.

I. The *NFP* and Rational Expectations

On October 6, 1979, the Federal Reserve adopted a new method for implementing monetary policy. Operating techniques in the 1970's had emphasized control of the federal funds rate to attain money growth targets. The new method focussed on the growth of bank reserves to accomplish desired monetary expansion and spotlighted the "money supply" relationship, where expansion of the monetary aggregates is linked to increments of bank reserves. The old procedure highlighted the "money demand" relationship, which linked the federal funds rate to growth in the demand for money.

The shift from interest rates to reserves as the focus of the Fed's operating approach was only one in a constellation of measures adopted between October 6, 1979 and mid-March, 1980. Further actions implemented on October 8, 11, and November 8, 1979 were designed to increase the reserves required against certain "managed liabilities" and to boost short-term interest rates, particularly on instruments used by banks for loanable funds.

This first set of measures did not bring the desired monetary growth by the end of 1979. In addition, because of concern over inflation and a federal government budget that was not sufficiently restrictive, the Carter Administration began to consider a range of new policies in early 1980, including credit controls. But even as the administration deliberated in February and early March 1980, the Federal Reserve aggressively drained bank reserves in a second shock to the financial markets. Short-term interest rates rose sharply under the *NFP*, by 3-1/2 to 6-1/2 percentage points between February 1 and March 17, 1980. And, the prime rate reached a record 20 percent during the week ending April 4, 1980.

Against this background, the Federal Reserve acted with the Carter Administration on March 14, 1980 to impose a third policy shock. Some of the central bank actions were based on the Credit Control Act of 1969 in an unprecedented use of credit controls during peacetime. Another round of higher reserve requirements was imposed on managed liabilities and new reserve requirements were placed on money market funds.

The three Fed policy shocks between October and mid-March fit the *RE* prescription for an effective policy. First, each change in policy was unanticipated; especially the shift to reserves from interest rates in the implementation of monetary policy. The subsequent draining of reserves in the winter and credit controls of March also were unexpected. Second, the break with tradition in the practice of monetary policy represented a change in structure that could not possibly be fully understood by market participants, thus interfering with the operation of the efficient markets hypothesis postulated by *RE*. The imposition of credit controls also was of the same genre. Third, with new feedback effects in place from the economy to Fed policy, economic agents could not correctly anticipate the course of action to be followed by the central bank. Indeed, the magnitudes or formulae by which reserves were to be changed were not made clear, creating confusion and making it difficult for economic agents to anticipate monetary events. Thus, with the major goal of the central bank to break the severe inflation in the *U.S.* economy, the *NFP* seems a logical action for a *RE* world.

II. The New Fed Policy: Analysis of Effects

Using the DRI Control forecast of August 24, 1980 as a baseline, a dynamic simulation

TABLE 1—EFFECTS OF REMOVING ALL ELEMENTS OF THE *NFP* FROM OCTOBER 6, 1979 TO MIDSUMMER 1980[a]
(AS SIMULATED IN THE DRI MODEL)

	1979	1980				1981				Years		
	IV	I	II	III	IV	I	II	III	IV	1980	1981	1982
Reserves:												
Required Reserves ($ bils.)	−0.05	−0.52	−1.87	−0.01	0.35	0.60	0.77	0.79	0.73	−0.54	0.72	0.35
Nonborrowed Reserves ($ bils.)	1.70	2.21	1.01	0.27	0.61	0.68	0.77	0.75	0.60	1.02	0.70	0.52
Interest Rates:												
Federal Funds (%)	−2.21	−4.22	−5.23	−1.62	−0.79	0.00	0.15	0.33	0.44	−2.96	0.23	0.16
Treasury Bills (%)	−1.64	−2.93	−3.23	−0.88	−0.50	0.01	0.05	0.20	0.25	−1.88	0.13	0.07
AAA-Equivalent Corporate Bonds New Issue Rate (%)	−0.15	−0.25	−0.16	0.00	0.17	0.20	0.42	0.43	0.54	−0.06	0.40	0.43
Monetary Aggregates:												
*M*1-A (% chg.)	1.45	2.67	3.74	1.75	1.65	0.87	0.24	−0.58	−0.89	2.45	−0.09	−0.95
*M*1-A (4 qtr. % chg.)	0.36	1.04	2.01	2.43	2.48	2.04	1.13	0.55	−0.09	1.99	0.91	−0.82
*M*1-B (% chg.)	1.39	2.58	3.60	1.66	1.55	0.79	0.15	−0.65	−0.93	2.34	−0.16	−0.91
*M*1-B (4 qtr. % chg.)	0.36	1.01	1.93	2.33	2.37	1.93	1.04	0.47	−0.16	1.91	0.82	−0.82
Loans:												
Commercial and Industrial ($ bils.)	−0.38	−2.56	−2.46	2.50	8.35	13.53	15.46	17.92	18.75	1.46	16.41	18.07
Consumer Credit Extended ($ bils.)	1.45	4.93	10.75	11.59	12.86	14.23	13.94	12.62	10.43	10.03	12.31	5.13
Economy:												
Real *GNP* (% chg.)	0.76	1.44	2.21	1.58	1.41	0.59	−0.60	−1.55	−1.91	1.66	−0.87	−1.07
Consumption (bils. '72 $'s)	0.70	2.82	7.06	11.76	14.67	15.63	14.45	11.67	9.04	9.08	12.70	4.26
Business Fixed Investment: (bils. '72 $'s)	0.44	2.07	5.17	8.62	11.50	12.96	12.55	10.54	8.05	6.84	11.03	3.12
Housing Starts (mils. units)	0.03	0.11	0.22	0.34	0.40	0.35	0.23	0.10	0.01	0.27	0.17	−0.07
Employment:												
Employment (mils. persons)	0.08	0.22	0.48	0.66	0.83	0.96	0.96	0.86	0.67	0.55	0.86	0.29
Unemployment Rate: (%)	−0.02	−0.07	−0.17	−0.31	−0.46	−0.60	−0.67	−0.66	−0.58	−0.25	−0.63	−0.23
Inflation:												
CPI Urban (%)	−0.04	−0.12	−0.02	0.17	0.29	0.28	0.19	0.18	0.08	0.08	0.18	0.06
Wholesale Price Index (%)	0.08	0.33	0.61	0.79	0.79	0.54	0.23	−0.01	−0.17	0.63	0.15	−0.27
GNP Deflator (%)	0.09	0.09	0.07	0.16	0.14	0.13	0.06	0.01	0.08	0.12	0.07	0.16

[a] Shown herein are differences between the historical dynamic simulation and the model simulation without the "temporary" and "permanent" elements of the *NFP*. The baseline was an historical dynamic solution with the residuals adjusted to produce actual history through 1980:2 and the DRI Control forecast of August 24, 1980 was used thereafter. Differences in % chg. or % are percentage points.

replicating history was created by feedback of residual errors prior to 1980:3. The *NFP* was then eliminated from history in a stepwise fashion. The "temporary" measures on reserve requirements and credit controls were removed by reducing reserve requirements on managed liabilities of member and nonmember banks to pre-October 6 levels; lowering the discount rate to the 11 percent of before October 6 and removing a three-point surcharge on member bank borrowing instituted March 17, 1980; raising the net return on money market funds in March to July by an amount equivalent to the 15 percent deposit requirement on increases of total money market fund assets above the outstanding of March 14; reducing required reserves by eliminating the special 15 percent deposit requirement on increases of outstanding credit over $2 million in credit cards, check credit overdraft plans, and unsecured loans; and adjusting consumer sentiment modestly upward to remove some of the deterioration in the outlook brought about by the *NFP*.

The "permanent" change in monetary policy of focussing on bank reserves rather than interest rates was removed in October 1979. Only small moves in the federal funds rate were permitted in response to deviations of monetary growth from target. Since monetary growth was above short-run targets only in 1979:4, the federal funds rate was raised 50 basis points and then permitted to decline by approximately the same amount as the economy fell into a recession. Once the recession began, larger declines occurred in keeping with past behavior. Beyond the third quarter, the model simulation was run with only the "permanent" change removed from the *NFP*. All changes in reserve requirements, the discount rate, and controls on credit were eliminated by midsummer and the policy inputs phased in to the baseline situation.

The results show a significant impact for the *NFP* on interest rates, monetary growth, the economy and employment (Table 1). Without the *NFP*, key short-term interest rates would have been lower by 164 to 523 basis points from 1979:4 to 1980:2, short-run growth in *M*1-*A* and *M*1-*B* higher by 1.4 to 3.7 percentage points, and the real economy stronger by 0.8 to 2.2 percentage points. Bond yields would have been somewhat lower at first, but then higher from the increased demand for liquidity under a better economy, eventually higher inflation rates, and lower unemployment. Housing starts were particularly impacted, down 220,000 units, at annual rates, in the second quarter of 1980. Consumer spending also was strongly affected, dropping a cumulated $10.6 billion, in real terms, from 1979:4 to 1980:4 and even more in 1981. Employment was lower by almost 1,000,000 persons by early 1981.

As simulated, the *NFP* was not necessary to slow the *U.S.* economy, with a sharp drop in real *GNP* likely during the second quarter in any case. The cumulated effects of rising interest rates during the previous two years, diminished purchasing power of households, and restriction induced by the interaction of high inflation rates and high taxes would have operated to bring a 7.4 percent drop in real *GNP* during 1980:2. The pattern for interest rates, monetary growth, loans, the economy, and employment would have been less variable without the *NFP*. A smoother path was indicated for all these variables relative to the baseline through the whole period of simulation. Little inflation benefit was indicated from the *NFP*, at most 0.8 percentage points for wholesale prices in 1980:3, although the transient nature for much of the *NFP* may have permitted little effect. Also, the DRI model specifications for inflation do not suggest a quick effect, but instead relatively long lags before inflation rates are altered.

III. Rational Expectations in Large-Scale Macro-Econometric Models

What does the model simulation reveal about the *NFP* and the ability of a large-scale macro-econometric model to analyze the implications of such a shock? The simulation suggests what *RE* propses, a decided impact on the economy and employment because of the unanticipated monetary policy shock. However, the effect on inflation was minimal, either in the short- or longer-run. The results raise the question of whether the extreme shock in monetary policy was necessary, although it could be argued that a model incorporating *RE* notions more fully might have come closer to tracking the actual drop in real *GNP* that occurred.

The DRI model results support the hypothesis that unanticipated changes in monetary policy significantly affect the *U.S.*

TABLE 2—ROLE OF EXPECTATIONS IN THE DRI MODEL—SOME MAJOR CATEGORIES

Impact	Expectations Variable(s)	Concept
Consumption:		
Furniture	Unanticipated Income, Permanent Income, Expected Inflation	*RE*, Extrapolative
Motor Vehicle & Parts	Unanticipated Income, Permanent Income, Expected Inflation	*RE*, Extrapolative
Other Durables	Unanticipated Income, Permanent Income, Expected Inflation	*RE*, Extrapolative
Clothing & Shoes	Unanticipated Income, Permanent Income, Expected Inflation	*RE*, Extrapolative
Food	Unanticipated Income, Permanent Income	*RE*, Extrapolative
Other Nondurables	Unanticipated Income, Permanent Income	*RE*, Extrapolative
Services—Other Household Operations	Permanent Income	Extrapolative
Services—Transportation	Unanticipated Income, Permanent Income	*RE*, Extrapolative
Investment:		
Nonresidential Equipment	Unanticipated Sales, Expected Debt Service Expected Capacity Utilization	*RE*, Extrapolative
Plant	Expected Debt Service Expected Capacity Utilization	Extrapolative
Housing-Single Family Starts	Expected Inflation-Median Sales Price of New Single Family Homes	Extrapolative
Interest Rates and Stock Prices:		
90-Day Treasury Bill Rate	Unanticipated Monetary Growth, Expected Inflation	*RE*, Extrapolative
New Issue Rate on AAA-Equivalent Corporate Bonds	Unanticipated Inflation, Expected Inflation, Expected Growth in the Monetary Base, Expected Stock Prices	*RE*, Extrapolative
S & P Index of 500 Common Stocks	Expected Growth in Earnings Per Share	Extrapolative
Wages:		
Hourly Earnings of Production Workers	"Short-Run" Expected Price Inflation, "Long-Run" Expected Price Inflation, Expected Unemployment	Extrapolative
Prices:		
Implicit *GNP* Deflators	Expected Unit Labor Costs, Expected Materials Prices, Expected Input Costs	Extrapolative
Producer Prices	Expected "Stage of Processing" or Input Costs	Extrapolative
Core Inflation	Expected Cost of Capital, Expected Productivity Growth	Extrapolative

economy, but still leave open the question of how much. If the DRI model were irrelevant for policy analysis, then simulations of the *NFP* should have produced no real difference with the baseline in any of the markets analyzed. Knowledge of the DRI model structure lessens the surprise of the results, since a considerable specification of expectations, both rational and extrapolative, exists in the model (Table 2). Unexpected shocks are modeled in some of the markets covered in the DRI model. These factors, along with the channels for the effects of monetary policy, produced the results in the simulation.

Extrapolative expectations always have had a large role in the DRI model structure. The notion of permanent income is modeled as an extrapolative, geometric lag on previous real disposable income. Expected inflation in the consumption block is specified as a permanent theory of expectations, using a second-order Pascal lag. In the equations for equipment and structures spending, expectations of sales, debt service, and capacity utilization are modeled as weighted averages of past values. A second-order Pascal lag formulation for the expected rate of inflation on home prices appears in the single family housing starts equation. Throughout the financial block of the DRI model, portfolio adjustments are postulated to depend upon expected own and alternative rates of returns. The expectation on inflation used in the long-term bond rate equation is a second-order Pascal lag of actual past rates of change in the price deflator for consumer goods, an extrapolative formulation. The inflation sector of the DRI model, formulated in a stage-of-processing approach, contains expectations on price inflation and unemployment for the determination of wages, expectations on unit labor costs, materials prices, and input costs for the implicit deflators, and expectations on the cost of capital and productivity growth in the determination of core inflation, all extrapolative.

Recent research in the DRI model has been focussing on "surprises" and "shocks," rather than fully extrapolative methods of incorporating expectations. By using deviations of actual from expected values to capture unanticipated shocks, some elements of *RE* appear. In the 1980 version of the DRI model, unanticipated income, defined as the difference between current real disposable income and permanent income, is highly significant throughout most of the equations for consumption. Since 1974, a "disappointment" variable has been used in the producers' durable equipment equation. Actual sales minus a geometric weighted average of past adjusted real final sales has had a highly significant negative coefficient.

$$
\begin{aligned}
(1)\quad 172 = {} & \underset{(2.65)}{12.87} - \underset{(.062)}{.083}\,KN72(-1) \\
& + \underset{(.077)}{.232}\,UEXP * KN72(-1) \\
& + PDL(2,6,FAR)\sum_{i=3}^{7} a(t-i)PQ/C(-i) \\
& + PDL(1,7,FAR)\sum_{j=1}^{7} b(t-j)DS(-j) \\
& - \underset{(.019)}{.096}\,(QEXP-Q) + \underset{(.909)}{3.07}\,DMY
\end{aligned}
$$

$\bar{R}^2 = .9965;\ DW = 1.685;$

$GLS(1958{:}1 \text{ to } 1979{:}4)$

$$\sum_{i=3}^{7} a(t-i) = \underset{(.003)}{.0136}\,; \quad \sum_{j=1}^{7} b(t-j) = \underset{(15.5)}{-47.9}$$

where 172 is producers' durable equipment spending in constant dollars; *KN72* is the net capital stock of equipment in constant dollars; *UEXP* is expected capacity utilization formed with a Koyck distributed lag; *P* is the implicit *GNP* deflator; *Q* is real final sales adjusted for pollution abatement expenditures; *C* is the rental price of capital; *DS* is the ratio of interest payments on debt to cash flow for nonfinancial corporations; *QEXP* is expected final sales, modeled by adaptive expectations; and *DMY* is a dummy variable for the investment buildup during the Vietnam War.

The surprise element in monetary policy, measured by the deviation of actual vs. expected growth in money, appears with only a small impact on the 90-day Treasury bill rate (equation (2)), but nevertheless is statistically significant. The effect of unanticipated inflation on bond yields is statistically significant in the equation for the new issue rate on top-quality corporate bonds (equation (3)). So far, notions on unexpected changes in monetary policy, employment, or demand have not been included in the inflation sector of the DRI model, but research is proceeding along these lines.

$$
\begin{aligned}
(2)\quad RB = {} & \underset{(.581)}{2.255} - \underset{(.344)}{1.760}\,FS \\
& + \underset{(.396)}{1.532}\,LOG(GNP72/N) \\
& + \underset{(1.26)}{6.169}\,LOG((PD/(P^*N)) \\
& /(PD(-1)/(P(-1)^*N(-1)))
\end{aligned}
$$

$$+\underset{(1.18)}{4.239}\,\mathit{LOG}((\mathit{PD}(-1)$$

$$/(P(-1)^{*}N(-1))/\mathit{PD}(-2)$$

$$/P(-2)^{*}N(-2)))+\underset{(.001)}{.005\,\mathit{BKLNS}}$$

$$/\mathit{DEPS}-\underset{(2.33)}{2.450}\,\mathit{FR}/\mathit{TR}+\underset{(.028)}{.731}\,\mathit{RCP}$$

$$+\mathit{PDL}(1,2,\mathit{FAR})\sum_{i=1}^{1}a(t-i)\dot{\mathit{PC}}$$

$$+\underset{(.008)}{.020}\,(\dot{M}1-A-\dot{M}1-\mathit{AEXP})$$

$$\bar{R}^2=.9924;\ \mathit{DW}=1.49;$$

$$\mathit{OLS}(1961{:}1\ \text{to}1979{:}4);\ \sum_{i=0}^{1}a(t-i)=\underset{(.019)}{.063}$$

$$(3)\quad \mathit{RL}=\underset{(1.35)}{-14.3}-\underset{(0.86)}{6.16}\,\mathit{LOG}((\mathit{NBR}$$

$$+\mathit{CURR}+\mathit{ADJ})/(P^{*}N))$$

$$+\mathit{PDL}(1,12,\mathit{FAR})\sum_{i=0}^{11}a(t-i)$$

$$\times\mathit{LOG}((\mathit{NBR})(-1)+\mathit{CURR}(-1)$$

$$+\mathit{ADJ}(-1))/(P^{*}N))(-1)/\mathit{NBR}(-2)$$

$$+\mathit{CURR}(-2)+\mathit{ADJ}(-2)/(P^{*}N)(-2))$$

$$+\underset{(.113)}{.951}\,\mathit{PCEXP}-\underset{(.010)}{.070}\,\mathit{PCEXP}^{*}\mathit{AVG.RU}$$

$$+\underset{(.077)}{.376}\mathit{RA}+\underset{(.021)}{.065}\,(\dot{\mathit{PC}}-\mathit{PCEXP})$$

$$+\underset{(.002)}{.0074}\,\mathit{JS\&PEXP}+\underset{(.592)}{6.598}\,\mathit{LOG}(\mathit{GNP72}/N)$$

$$+\underset{(3.80)}{6.345}\,\mathit{LOG}(\mathit{CB}/(P^{*}N))$$

$$/(\mathit{CB}(-1)/P(-1)^{*}N)))$$

$$+\underset{(1.70)}{2.038}\,\mathit{LOG}(\mathit{PD}/P^{*}N))/(\mathit{PD}(-1)/)$$

$$\times(P(-1)^{*}N(-1)))+\underset{(.074)}{.214}\,\mathit{RMDIFF}$$

$$+\underset{(.146)}{.295}\,\mathit{DMYVIET}$$

$$\bar{R}^2=.9873;\ \mathit{DW}=1.99;$$

$$\mathit{OLS}(1954{:}1\ \text{to}\ 1979{:}4);$$

$$\sum_{i=0}^{11}a(t-i)=\underset{(9.81)}{33.42}$$

where *RB* is the Treasury bill rate; *FS* is foreign holdings of short-term *U.S.* government debt relative to outstanding; *GNP72* is real *GNP*; *N* is total population; *PD* is private debt outstanding; *P* is the implicit *GNP* deflator; *BKLNS* is total bank loans; *DEPS* is bank deposits; *FR* is free reserves and *TR* total reserves; *RCP* is the commercial paper rate; $\dot{PC}$ is the annual rate of change in the consumer goods deflator; $\dot{M}1$-*A* is the growth in narrow money and $\dot{M}1$-*AEXP* is the expected growth; *RL* is the average yield on new issues of high-grade corporate bonds; *NBR* is nonborrowed reserves; *CURR* is currency; *ADJ* is the reserve adjustment for changes in reserve requirements; *PCEXP* is the expected rate of inflation on the consumption goods deflator, formed on the basis of a second order Pascal lag; *RU* is the unemployment rate; *JS&PEXP* is expected growth in the Standard & Poor 500 Common stock price index; *CB* is outstanding nonfinancial corporate bonds; *RA* is the yeild AAA state and local government bonds; and *RMDIFF* and *DMYVIET* are dummy variables.

It is an open question if and how much bias exists in policy simulations with extrapolative expectations specifications or elements of rational expectations such as are included here. Certainly a model with a pervasive presence of expectations does not resemble the construct criticized by *RE*. A major challenge to the testing of a more pure form of *RE* in large-scale macro-econometric models lies in defining and specifying the rationally determined expected values of the variables about which deviations occur. If the model itself provided the correct structure of the economy and perfect information existed, then the projections of the model would turn out to be the rational expectations and deviations in actual performance about these predictions would have a major effect on both the model parameters and realized values of the forecasted variables. But, so far, the technology of large-scale modelling does not permit forecasts of the model to be used as expectations, then realized values about these expectations to be included as inputs.

IV. Concluding Comments

Are arguments against the use of macroeconometric models in policy analysis valid or are the results from model simulations robust with respect to the *RE* criticism? The results of simulating the *NFP* in the DRI model, with a genesis as if it were born out of *RE*, show a significant short-run impact on the economy. With a wealth of expectations modeled, although not as unbiased measures of expectations formation, these results suggest robustness for the large-scale model approach with respect to *RE* criticisms. Further, because model proprietors can incorporate "nonmodel" effects in simulations, the model approach offers a viable method for analyzing the impacts of policy. Effects on structural parameters or the behavior of unexpected changes in policy can be incorporated, or the range of results ascertained in sensitivity analysis, given the levers for interaction through constant add adjustments and exogenous variables.

Second, the arguments against the use of large-scale macro-econometric models in policy analysis are based on *RE* assumptions and hypotheses about the real world. But the notion that economic agents perceive and use the correct structure in forming expectations is hardly plausible and continuous equilibria do not seem possible. A homogeneity of views across decision makers on the determinants of the economy, interest rates, and responses to policy is quite unlikely. The instantaneous equilibrium of *RE* is anathema to the real world since disequilibria characterize quarter-by-quarter adjustments.

Third, not all economic markets are characterized by instantaneous, perfect information on those activities relevant to decisions. While the flow of information and reactions of economic agents is near perfect in the financial markets, reactions in labor markets are much less so, with contracts making for sticky responses. Other markets are not likely to be efficient, at least over periods so short as a quarter.

Fourth, as has been pointed out by Kenneth Arrow, in *RE* "economic agents are required to be superior statisticians, capable of analyzing the future general equilibrium of the economy" (p. 160). This is a difficult pill to swallow since not even elaborate, detailed specifications of economic processes that incorporate large bodies of data have achieved a sufficient degree of success in describing the economy.

Fifth, convergence to rational expectations equilibria may be slow or perhaps never occur. Policy changes and changes in structure occur frequently enough so that economic agents may never incorporate in reactions the expectations on policy that will fully anticipate the impacts.

Sixth, it is not clear that the errors potentially introduced by using a formulation of expectations different from *RE* are large enough to justify the charges levied at econometric models by the *RE* theorists. Estimates, theoretical or empirical, of the bias from using extrapolative as opposed to rational expectations mechanisms are not yet available. Impacts on model behavior of deviations between actual and extrapolated expectations could approximate the true effects closely enough for useful policy analysis even if the expectation is not unbiased.

Finally, the burden of proof for *RE* hypotheses really should rest with the proponents of *RE*. Though the logic of the *RE* arguments, given the assumptions, is incontrovertible, it is an empirical question whether significant variation occurs in structural coefficients when policy is changed. Methods for testing such a notion exist, for example, "recursive residuals" where additional periods are added to a regression during a suspected change of structure to see how the coefficients respond. In other areas of economics, untested theories are subjected to the discipline of empirical tests and estimation. Why should *RE* be different?

REFERENCES

K. J. Arrow, "The Future and the Present in Economic Life," *Econ. Inquiry*, Apr. 1978, *16*, 157–69.

Economies of Scope

By JOHN C. PANZAR AND ROBERT D. WILLIG*

Several years ago we coined the term "economies of scope" to describe a basic and intuitively appealing property of production: cost savings which result from the *scope* (rather than the *scale*) of the enterprise. There are economies of scope where it is less costly to combine two or more product lines in one firm than to produce them separately. While the concept itself is not completely novel, especially since multiproduct firms are the rule rather than the exception in our economy, we have attempted to make this terminology precise, both in common parlance and in theoretical analyses. Although our definition of economies of scope does not correspond exactly to joint production in the Marshallian sense, we show, in Section I, that it precisely characterizes the conditions which lead to the formation of multiproduct firms in perfectly competitive markets.

However, this formal equivalence, in and of itself, provides only limited insight into the tangible forces which make it feasible and profitable to form multiproduct enterprises. In the classic works of John Clark, Eli Clemens, and others, it was suggested that the origins of the multiproduct firm spring from the opportunity to exploit some type of excess capacity. This notion seems to imply that, when there are economies of scope, there exists some input (if only a factory building) which is shared by two or more product lines without complete congestion. Section II examines this issue in general and with a *micro* model of the technology which explicitly posits the presence of a *sharable*, "quasi-public" input.

Whenever the costs of providing the services of the sharable input to two or more product lines are *subadditive* (i.e., less than the total costs of providing these services for each product line separately), the multiproduct cost function exhibits economies of scope. Nevertheless, the precise nature of such economies of sharing has profound implications for how the boundaries and scope of the firm may be affected by transactions costs, market failures, and other Coasian considerations. In general, the presence of a quasi-public input mandates the existence of multiproduct firms *at some level* in the chain of production, even if there are no imperfections in the market for the input itself. However, if the shared-input services are undifferentiated among end uses, multiproduct firms result from the failure of the market to sustain efficient vertical *disintegration*, as discussed by George Stigler.

I. Economies of Scope and Multiproduct Firms

Let $N=\{1,2,\ldots,n\}$ denote the set of products under study, with respective quantities $y=(y_1,\ldots,y_n)$. Let y_S denote the n vector whose elements are set equal to those of y for $i\in S\subset N$ and 0 for $i\notin S$. The function $C(y_S,w)$ denotes the cost of producing only the products in the subset S, at the quantities indicated by the vector y. Here, $C(y,w)$ is the usual multiproduct minimum cost function, and w is the vector (usually suppressed) of factor prices. We are now able to provide a formal definition of our fundamental concept.

Definition: Let $T=\{T_1,\ldots,T_l\}$ denote a nontrivial partition of $S\subseteq N$. That is, $\cup_i T_i=S, T_i\cap T_j=\varnothing$ for $i\neq j$, $T_i\neq\varnothing$, and $l>1$. There are *economies of scope* at y_S and at factor prices w with respect to the partition

*Bell Telephone Laboratories, Inc. and Princeton University, respectively. The views expressed are our own and do not necessarily reflect those of Bell Laboratories or the Bell System. We are indebted to David Teece for helpful comments.

T if

$$(1)\qquad \sum_{i=1}^{l} C(y_{T_i}, w) > C(y_S, w)$$

There are said to be *weak economies of scope* if the inequality in (1) is weak rather than strict, and *diseconomies of scope* if the inequality is reversed.

Economies of scope can be immediately related to the structure of firms in classical competitive equilibrium.

PROPOSITION 1: (i) *Economies of scope at prevailing factor prices with respect to the specialization partition* $T=\{1,2,\dots,n\}$ *are sufficient for the existence of multiproduct firms in multiproduct competitive equilibrium, and* (ii) *weak economies of scope with respect to all partitions of at least one nontrivial* $S\subseteq N$ *are necessary for such existence of a multiproduct firm.*

PROOF:

(i) Suppose the contrary; that is, that equilibrium involves only single product firms, each earning zero profits. Then, a merger of n price-taking firms, each specializing in a different product, would lower their total costs and yield positive profit. Hence, firms would have a profit and the hypothesized industry configuration could not be a competitive equilibrium.

(ii) Suppose there was a multiproduct firm producing the product set S in competitive equilibrium, and there were diseconomies of scope with respect to the nontrivial partition of S $\{T_1,\dots,T_l\}$. Then a firm producing only the products in T_j, for some j, could earn positive profit at the equilibrium prices p, since

$$0=\sum_{S} p_i y_i - C(y_S) < \sum_{j=1}^{l}\left[\sum_{T_j} p_i y_i - C(y_{T_j})\right]$$

Thus, if there did not exist at least one subset of N, containing two or more products, which was invulnerable to such fragmentation, no multiproduct firms could exist in competitive equilibrium.

Having established the consequences of the presence or absence of economies of scope for the structure of firms in competitive equilibrium, we now turn to an examination of input sharing, an important source of these economies.

II. Shared Inputs and Economies of Scope

It is intuitively appealing to link economies of scope to the existence of sharable inputs; that is, inputs which, once procured for the production of one output, would be also available (either wholly or in part) to aid in the production of other outputs. Examples of sharable inputs might include elements of productive capacity (such as electric power generators or transmission facilities) usable at different times for different outputs, indivisible equipment (or just a factory building) usable for more than one manufacturing process, heat sources only partially depleted by their primary uses, human capital applicable to the production of more than one output, or inputs (such as sheep) which inevitably offer by-products (such as mutton) from their primary production (such as wool).

Two categories of issues arise from consideration of the links between sharable inputs and economies of scope. First are the issues of the definition of a sharable input, of how one can be identified from a description of the technology, and of whether economies of scope are equivalent to the existence of a sharable input. Second are the issues concerning the vertical structures of firms and markets for the services of sharable inputs. We defer for Section III our remarks on the latter set of issues.

It would seem most natural to *define an input as sharable between the productions of product sets S and T if the joint production of these outputs enables some of the input to be conserved, vis-à-vis separate production, while the utilizations of all other inputs were not expanded.* With this definition it can be shown, under several standard regularity conditions, that *there are economies of scope between product sets S and T for all positive factor-price vectors if and only if there exist inputs sharable between them.* (Unfortunately, space permits neither a precise statement nor a proof of this result. Both are available from the authors. The proof entails a new strong form of Hirofumi Uzawa's theorem on the duality between cost functions and technology sets.)

Although this is a strong result that demonstrates an equivalence between economies of scope and the existence of sharable inputs, it leaves open some related questions that merit further research. The result does not specify how the identities of the sharable inputs can be recovered from either the cost function or the dual technology set. In fact, when inputs are readily substitutable within the processes utilized to produce product sets S and T, the identities of the sharable inputs, as defined above, may be ambiguous. Also, the result does not apply to circumstances in which the presence of scope economies depends on the relative levels of factor prices. This can be a serious omission if the sharable inputs are excluded from the cost-efficient input bundles at some factor prices. To circumvent the current limitations of our abstract duality approach to economies of scope, and to pave the way for our discussion of firm and market structure, we now specify a more concrete model of input sharing.

The micro model of the technology which we study involves n otherwise independent production processes which are able to share the services of some productive inputs. For ease of exposition, we assume that there is only one such input, K, called "capital." Then the multiproduct minimum-cost function, which embodies the least costly way of producing y_S, results from solving the program:

$$(2)\quad C(y_S) \equiv \min_{k} \sum_{i \in S} V^i(y_i, k_i) + \psi(k, \beta)$$

where V^i represents the minimum variable cost of producing the output y_i using k_i units of capital *services*. The capital service cost function $\psi(k, \beta)$ represents the cost of acquiring the requisite vector k of capital services, where β represents relevant factor prices. Finally, let $k^*(y_S)$ denote the *argmin* of (2), the cost-minimal vector of capital services required for the production of y_S. We simply assume here that $k^*(y)_i > 0$ for $y_i > 0$ at prevailing input prices, while $k^*(y)_i = 0$ for $y_i = 0$.

We shall focus on cases in which ψ is *strictly* subadditive in k. Another way to describe such situations is to say that K is a *quasi-public* input, since its services can be shared by two or more product lines without complete congestion. The most extreme form occurs when, as in the simple peak load pricing model, capital is a pure public input; i.e., $\psi(k, \beta) = \beta K = \beta \max_i k_i$. On the other hand, when capital is a competitively marketed pure private input, $\psi(k, \beta) = \beta K = \beta \Sigma_i k_i$, which is only weakly subadditive in k. Intermediate cases also fit well into this framework, hence the term quasi-public input. We are now in a position to state:

PROPOSITION 2: *For any nontrivial partition of N, there are economies (diseconomies) of scope if and only if ψ is strictly subadditive (superadditive) in the relevant range.*

PROOF:

Let $\{T_1, \ldots, T_l\}$ be a nontrivial partition of N, and let

$$\hat{k} = \sum_j k^*(y_{T_j})$$

It follows from (2) and the definition of $k^*(\cdot)$ that

$$(3)\quad \sum_N V^i(y_i, k_i^*(y)) + \psi(k^*(y)) = C(y) \leqslant \sum_N V^i(y_i, \hat{k}_i) + \psi(\hat{k})$$

$$(4)\quad \sum_{T_j} V^i(y_i, k_i^*(y)) + \psi(k^*(y)_{T_j}) \geqslant C(y_{T_j}) = \sum_{T_j} V^i(y_i, k_i^*(y_{T_j})) + \psi(k^*(y_{T_j}))$$

Summing (4) over j and subtracting from (3) yields

$$(5)\quad \psi(k^*(y)) - \sum_j \psi(k^*(y)_{T_j}) \leqslant C(y) - \sum_j C(y_{T_j}) \leqslant \psi(\hat{k}) - \sum_j \psi(k^*(y_{T_j}))$$

The conclusions follow since the leftmost (rightmost) term in (5) is positive (negative) if and only if ψ is strictly superadditive (subadditive) over the relevant range.

Thus, the micro model we have constructed illustrates the equivalence between the existence of quasi-public, sharable, inputs and economies of scope.

III. The Scope of the Firm

The micro model of production just presented can be utilized to shed light on issues concerning the structures of the firm and the input markets in which it participates. Proposition 2 shows that there are economies of scope in the operations of a firm that produces its own stream of services from a quasi-public input and utilizes them for the production of more than one output. And Proposition 1 shows that there must be multiproduct firms in competitive equilibrium, where economies of scope are prevalent.

However, the model from which Proposition 2 is derived *presumes* that the services of the quasi-public input are self-produced. If, instead, these services could be efficiently allocated by a market, then both the economies of scope and the need for horizontal integration over the final products in the set N would disappear. Hence, in this limited framework, questions on the scope of the firm devolve to questions on the efficient marketability of the services of the quasi-public input.

Efficient market allocation of these services to vertically disintegrated specialty firms would require that firms producing y_i face an input price equal to marginal cost $\partial\psi/\partial k_i$, if ψ is differentiable. Hence, when the different end uses incur different marginal costs of providing the shared-input services, these services must, for efficiency, bear user differentiated prices. (Of course, this is also true of the Lindahl prices associated with a pure-public input.) On the other hand, if competitive markets can sustain such prices, then it may be said that the effect of scope economies is but "pushed back" one level to multiproduct suppliers of shared-input services.

Consider, for example, the classic peak-load pricing model with but two time periods. A power plant's generating *capacity* can clearly be viewed as a pure public input. The "downstream industries" in this case would be composed of firms purchasing fuel and contracting for capacity services, either during the day or the night, in order to market day or night electricity to final consumers. Given a competitive supply of generating capacity, there is no positive incentive for such "firms" to provide both products; that is, serve both times of day. However, the upstream industry is comprised of multi-output firms, each of which must, in general, sell capacity services to both downstream markets and charge them *different* prices (peak vs. off peak) in competitive equilibrium.

On the other hand, arbitrage or the difficulty of labelling purchasers may render infeasible user-differentiated market prices for the shared-input services. In such cases, self-production of the shared-input services, economies of scope, and multi-output production of the final goods are all inevitable in competitive equilibrium.

Additional factors influence the scopes of the firms, even if all types of specialized firms incur equal marginal costs of shared-input services. In our model, this case is equivalent to $\psi(k)\equiv F(\Sigma k_i)$. At one extreme, $\psi(k,\beta)\equiv\beta\Sigma k_i$, the input is completely private, and its competitive allocation to specialty firms is efficient. The same result holds if the average costs of providing the input services, $F(z)/z$, reach a minimum at a scale that is small relative to the total derived demand for them. At the other extreme, global scale economies in the provision of $z=\Sigma k_i$ can make financially infeasible marginal cost pricing of the input services, and may thus impart both economies of scope and natural monopoly to the final output market.

Finally, any type of market transactions costs and any of the many other causes of vertical integration can lead self-production of shared-input services to be the efficient (or just profitable) alternative. In such instances, Proposition 2 delineates the implications for the scope of the firm.

IV. Concluding Comments

When the multiproduct cost function summarizes *both* the production *and* organizational costs of operating a firm, economies of scope is the precise condition required for the emergence of multiproduct firms in a competitive environment. However, as David Teece has argued, application of the concept to production costs alone may mask the economic factors that influence the scope of firms. Our quasi-public input model shows that multiproduct firms (at some stage in the production process) are inevitable when the marginal costs of providing the services of the shared input vary across product lines. In general, and especially when said services are undifferentiated across uses, the scope and the vertical structure of a firm are inextricably related. We hope that these insights will provide some guidance and impetus for future research on the micro-organizational structure of multiproduct firms.

REFERENCES

John M. Clark, *Studies in the Economics of Overhead Costs*, Chicago 1923.

E. Clemens, "Price Discrimination and the Multiproduct Firm," in Richard Heflebower and George Stocking, eds. *AEA Readings in Industrial Organization and Public Policy*, Homewood 1958, 262–76.

J. Panzar and R. Willig, "Economies of Scale and Economies of Scope in Multi-Output Production," econ. disc. paper no. 33, Bell Laboratories 1975.

G. Stigler, "The Division of Labor is Limited by the Extent of the Market," *J. Polit. Econ.*, June 1951, *59*, 185–93.

D. Teece, "Economies of Scope and the Scope of the Enterprise," *J. Economic Behavior, Organization*, forthcoming.

H. Uzawa, "Duality Principles in the Theory of Cost and Production," *Int. Econ. Rev.*, May 1964, *5*, 216–20.

Sustainability and the Entry Process

By KENNETH C. BASEMAN*

The work of John Panzar and Robert Willig, and Gerald Faulhaber has established that a natural monopoly cannot necessarily survive under free entry. Even if production of any given industry output is always cheapest if undertaken by a single firm, entry may occur in the industry. The resultant set of products and prices may generate lower levels of economic welfare than the pre-entry equilibrium. Thus the "market test" provided by a free entry policy in regulated, natural monopoly markets may be an ineffective test of the performance of the incumbent firm.

Before accepting this conclusion, some further questions must be answered. First, in viewing entry as providing a market test of a putative natural monopoly, one must have in mind some assumption about the incumbent firm's reaction to entry, as well as the restraints, if any, which the regulator imposes on the incumbent's response to entry. It is quite possible that, under one set of rules governing the incumbent's behavior, entry will provide a poor test of the incumbent's performance, while under another set of rules free entry will provide a good test. Second, the desirability of free entry in natural monopoly markets depends on the incumbent's behavior absent an entry threat. Second best welfare maximization and profit maximization are two possible specifications of the incumbent's behavior, and they have different implications for the desirability of free entry.

I. Principles for a Good Market Test

There are two principles against which I propose to judge whether a particular set of rules governing the entry process provides a good test of the performance of an existing firm:

1) A Ramsey firm (defined here to be a firm which chooses, subject to a break-even constraint, the welfare-maximizing set of products, prices, and outputs) will be sustainable, that is, it will survive the market test and entry will not occur.

2) If entry occurs, welfare will increase over the pre-entry levels. This principle assures that it is not merely an initial failure by the incumbent to produce at the welfare optimum which brings forth entry. Rather, the successful entrant must generate an increase in welfare.

The reactions to entry the incumbent is allowed in the Faulhaber and Panzar-Willig models are severely limited. The incumbent firm is not allowed to change its price(s) when faced with entry in one of its markets. In that environment, they show that the second principle is violated.

William Baumol, Elizabeth Bailey, and Willig, given the same restrictions on the reactions of the incumbent to entry into one or more of its markets, provide a set of assumptions about costs and demands under which a Ramsey firm will be sustainable.

My forthcoming article shows that allowing the incumbent firm to reduce price in a profit-maximizing (nonpredatory) manner will satisfy the second principle in the case where (i) the incumbent firm serves one market and the entrant offers the same product, (ii) the incumbent's average nonsunk costs are less than the entrant's average total cost,[1] and (iii) the incumbent's price equals average cost before entry occurs. That result does not extend to a multiple product environment. In that case profit-maximizing price reactions by an incumbent to entrant duplicating one of its

*Economic Policy Office, Antitrust Division, U.S. Justice Department. The views expressed are my own, not necessarily those of the Antitrust Division. I thank Robert Reynolds for his helpful comments and theoretical assistance. I am solely responsible for any errors.

[1] In that discussion I implicitly treated the fixed costs as being entirely sunk.

products will not result in the satisfaction of the second principle—successful entry will not necessarily improve welfare. Further, Panzar argued, correctly, that complete price flexibility would not necessarily increase the prospects for sustainability. A natural monopoly which is sustainable if no price reaction is allowed may be unsustainable if it is allowed price flexibility.

The only analysis of an entrant providing an imperfect substitute for the product(s) of a regulated monopolist has been provided by Ronald Braeutigam. He argued that if firms set quantities, not prices, successful entry by a firm with a new product will not necessarily increase welfare. Robert Reynolds considered the product choices of a nonregulated, profit-maximizing natural monopolist. For the specific demand structure he analyzed, he found that the natural monopolist would be sustainable.

II. Noncooperative Price Setting with Entry by Firms Offering Differentiated Products

A principal remaining gap in this literature is analysis of noncooperative price setting by the firms when the entrant offers a differentiated product. Some of my recent work has been directed at this problem. The main finding of my initial modelling efforts is that such price competition does not satisfy the two principles for an effective market test.

The entrant considers offering a new product at a particular price. The incumbent firm, subject to the profit constraint, charges the profit-maximizing price given the new demand curve it faces. The entrant expects the monopolist to respond to entry in this profit-maximizing manner, and enters if his expected profits, given incumbent firm's expected reaction, are greater than zero. Thus the game is played in prices and, due to the asymmetry between the entrant and the incumbent, the entrant is viewed as a Stackleberg leader.

A Ramsey firm is not necessarily sustainable. This proposition will be established by construction of an example of an unsustainable Ramsey firm. In constructing the example, I will confine the analysis to the simplest case, where a single product, product 1, is offered by the incumbent firm and a substitute, product 2, is offered by an entrant. Superscript bars and "hats" will refer respectively to pre-entry and post-entry values of price (p) and quantity (q). The cost functions feature constant variable (and marginal) costs, v_1 and v_2, which are the same for both firms. The industry is a natural monopoly because the fixed (and sunk) costs are lower for a firm offering both products than for two firms: $F(2) > F(1,2) - F(1) \equiv AF(2;1)$. That is, the fixed costs incurred by the entrant, $F(2)$, exceed the additional fixed costs, $AF(2;1)$, the incumbent would incur in adding product 2, given production of product 1.

The proposition that a Ramsey firm is not necessarily sustainable will be established in three steps. A: If the incumbent must hold p_1 constant at $\bar{p}_1$ when product 2 is added, then entry may be feasible even though the incumbent would not add product 2. B: If we allow the incumbent to change price post-entry, he may well choose to increase price. Then entry profitable at $\bar{p}_1$ will certainly remain profitable at higher prices for product 1. C: If entry is profitable at $\bar{p}_1$, it is not necessarily true that it would have been Ramsey optimal, with a different p_1, for the incumbent to have offered both products initially.

A: If the products are substitutes and the incumbent adds product 2 to his product set, holding p_1 constant at $\bar{p}_1$, then the incumbent will incur a loss in contribution to fixed costs, call it C, in market 1. That situation is depicted in Figure 1. Suppose the entrant's price $\hat{p}_2$ is equal to the price which would have maximized the total revenue contribution from both markets if the incumbent had offered service 2 while holding $\bar{p}_1$ fixed. The revenues the incumbent earns from market 2 must cover the variable costs of product 2, AF, and C if the incumbent is to offer the product and still break even. The entrant, on the other hand, must cover only the variable costs plus $F(2)$, and will be unconcerned about the effects of entry on the revenues earned in market 1. Now we know $F(2) > AF$, but the difference can be arbitrarily small. In Figure 2, a situa-

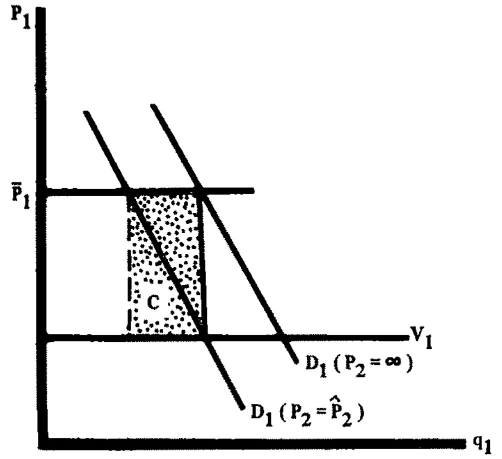


FIGURE 1

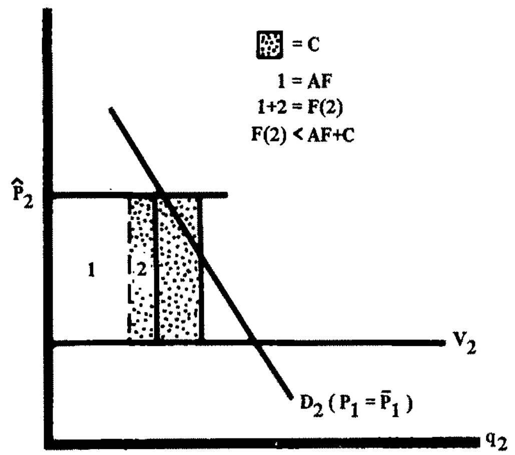


FIGURE 2

tion is depicted where revenues in excess of variable costs cover $F(2)$, but not $AF+C$, when service 2 is offered at $\hat{p}_2$.[2] Thus entry is feasible, although the incumbent could not profitably offer the product while holding $\bar{p}_1$ fixed.[3]

B: If entry in market 2 is feasible at $\bar{p}_1$, it is also feasible for higher prices of product 1. When freed from the post-entry price constraint, the incumbent may choose to raise its price. For example, this will occur if the profit function for service 1 is a concave function of p_1 for given values of p_2 and it is feasible to break even after entry occurs.

C: Is the hypothesized profitability of entry given p_1 inconsistent with an initial position of Ramsey optimality? We start at $(\bar{p}_1, \hat{p}_2)$. Could the incumbent firm, free to restructure its prices, have necessarily picked a new set of prices which would yield greater welfare than producing only product 1 at $\bar{p}_1$? The answer is no. An example will serve to establish the point. Suppose that at the price ratio $\bar{p}_1/\hat{p}_2$ there exists a group of consumers for whom the products are very close substitutes. When they switch from product 1 to product 2, their total consumers' surplus increases only an arbitrarily small amount. The remaining consumers do not view the products as good substitutes at all, so additional reductions in p_2 relative to p_1 will generate only arbitrarily small increases in the consumption of product 2. The situation is depicted in Figure 3. The entrant can just barely recover its cost at $\hat{p}_2$. But the incumbent will not necessarily be able to offer both products at higher welfare levels.

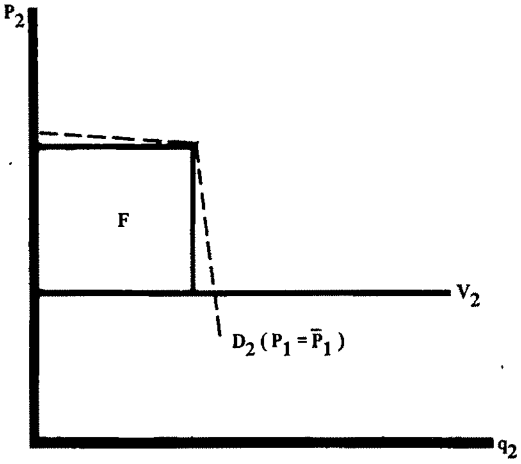


FIGURE 3

Consider the change in consumer's surplus resulting from a small increase in p_1, which, at the original price pair, can be

[2] This is a sufficient condition, but not a necessary one, for profitable entry given $\bar{p}_1$.

[3] An implication is that if the incumbent chooses to add products based on a "burden" test, where consumers in other markets cannot be charged higher prices as the result of adding a new product, a service not offered by the incumbent because it fails the burden test may be offered by an entrant.

approximated by

$$(1)\quad \Delta CS=(p_1-v_1)\Delta q_1+(p_2-v_2)\Delta q_2$$

The first term is negative. The second term is positive, but small because Δq_2 is arbitrarily small. The welfare effect of the price change given Δq_2 small enough, will be negative.

The incumbent, in this situation, might be able to satisfy the break-even constraint at some price pair $(\tilde{p}_1, \tilde{p}_2)$, where $\tilde{p}_1 > \bar{p}_1$ and $\tilde{p}_2 > \bar{p}_2$. The incumbent can extract additional revenues from market 2 by increasing p_2 along with p_1, thereby maintaining the price ratio which triggers the demand for service 2.[4] However, the incremental surplus from providing service 2 remains very small, and can easily be less than the losses in consumer surplus of those consumers who must pay higher prices for product 1. I have shown that a Ramsey firm may be unsustainable, even given post-entry price flexibility. By a similar argument, it can be shown that the second principle for a good market test also fails to hold here.

III. Sustainability and Regulation of the Entry Process

Sustainability analysis has established that entry into natural monopoly markets may not always be desirable. This conclusion raises (at least) two interesting questions. First, how do we go about separating markets where entry is desirable from those where it may have adverse consequences? Second, does the new economics of entry in natural monopoly markets imply that regulators should reassess the reasoning behind their recent pro-entry, procompetition decisions?

With regard to the first question, some further theoretical work appears promising. The new "contestability" analysis (see Bailey for a discussion) offers an explanation of why entry may be desirable in some natural monopoly markets when sunk costs are small.

Another avenue is to examine the behavior of the incumbent firm absent the threat of entry. If the incumbent will find it in its interest to choose an equilibrium set of outputs and prices which generates lower welfare than the free entry equilibrium, then free entry is a desirable policy. Entry will make things better, even though it will not necessarily result in or sustain a second best welfare optimum.

Along these lines, Robert Reynolds and I are currently examining the behavior of a rate-of-return constrained monopolist in a one-dimensional model of product differentiation. Absent an entry threat, profit maximization by such a firm will result in more products than is socially optimal. Free entry will push the incumbent firm in the right direction; the firm, in order to deter entry, will choose to offer fewer products. The interesting remaining question is whether there will be "overshooting." If the number of products under free entry also exceeds the optimum, then free entry is unambiguously desirable.

If unsustainability is not a problem, or is a significantly smaller problem, when the incumbent is a profit maximizer than when the incumbent is a Ramsey firm, then there are interesting prospects for empirical work. The two competing behavioral hypotheses generate different empirically testable predictions.

As for the second question, it does not appear that the sustainability literature has provided new areas of inquiry, previously overlooked by regulators, which would be cause for reconsideration of the recent moves to open regulated markets to competition. Regulators have not believed that allowing competition is a risk-free policy. Moreover, they appear to have asked themselves exactly the questions the sustainability literature says they should: what are the effects of competition in some markets on the price and availability of products not expected to be offered by the entrants? When allowing entry, they have generally concluded that these expected price effects were small, or were balanced by other con-

[4] A profit-maximizing, rate-of-return regulated monopolist could find it attractive to offer both products, since offering the second product would increase both actual and allowed revenues.

siderations such as improved cost efficiency, pricing rationality, or innovation.

REFERENCES

E. E. Bailey, "Contestability and the Design of Public Policy for Monopoly," discus. paper, June 1980.

K. C. Baseman, "Open Entry and Cross-Subsidy in Regulated Markets," in Gary Fromm, ed., *Economics of Public Regulation*, Cambridge, Mass., forthcoming.

W. Baumol, E. Bailey, and R. Willig, "Weak Invisible Hand Theorems in the Sustainability of Multiproduct Monopoly," *Amer. Econ. Rev.*, June 1977, *67*, 350–65.

R. R. Braeutigam, "The Regulation of Multiproduct Firms: Decisions on Entry and Rate Structure," unpublished doctoral dissertation, Stanford Univ. 1976.

G. Faulhaber, "Increasing Returns to Scale: Optimality and Equilibrium," unpublished doctoral dissertation, Princeton Univ. 1975.

J. Panzar, "Comment," in Gary Fromm, ed., *Economics of Public Regulation*, Cambridge, Mass., forthcoming.

______ **and R. Willig,** "Free Entry and the Sustainability of Natural Monopoly," *Bell J. Econ.*, Spring 1977, *8*, 1–22.

R. Reynolds, "Product Selection and Entry Reaction," discussion paper, Dept. of Justice, July 1978.

On the Political Sustainability of Taxes

By JANUSZ A. ORDOVER AND ANDREW SCHOTTER*

Economic analysis of commodity taxation proceeds on the assumption that such taxes are set by a social planner who is empowered to maximize some social welfare function (*SWF*) subject to a predetermined budgetary requirement (see A. B. Atkinson and J. E. Stiglitz). In Western democracies, however, excise taxes are most often determined within a political process. Hence, it is important to inquire whether the commodity taxes that result from this political process differ from the constrained welfare-maximizing (Ramsey-optimal) taxes which would have been selected by the omnipotent social planner.[1]

To study this question we posit that social planners aim to minimize the excess burden of commodity taxes. This assumption allows us to abstract from equity considerations and provides us with a convenient benchmark vector of Ramsey-optimal commodity taxes t^*. We construct three different models of the political process. In the context of each of these models, we investigate whether a set of politicians who are competing for voters in a general election by proposing tax vectors, will find the Ramsey-optimal tax vector among the set of sustainable or equilibrium tax vectors. We demonstrate that, in general, for most economies and for a variety voting models, the *politically sustainable* tax vector t^{**} diverges from t^*, where a politically sustainable tax vector is defined as a tax vector which, if set by a politician, can either guarantee his election (if no two politicians can set the same tax vector) or characterize the equilibrium of the political race (if identical tax proposals are allowed).[2]

The reason for this divergence can be explained by the fact that whereas the Ramsey-optimal taxes reflect the intensity of consumers' preferences over tax vectors, as represented by the deadweight loss of the tax vector, politically sustainable tax vectors are determined primarily by the consumer-voters' ordinal preferences. This point emerges clearly in Section II where we study voting models which differ in the extent to which the intensity of preferences over the feasible tax vectors determines the outcome of elections. In the *pure democracy* voting model of Section IIA the only determinant of the election's outcome are the candidates tax proposals and the voters' ordinal preferences over these proposals. Here the winning tax vector is that which is preferred by the requisite majority of voters. In this model, we find that very restrictive and unrealistic assumptions on the source of voters' heterogeneity are needed to ensure the equivalence between Ramsey-optimal and politically sustainable taxes.

The pure democracy model is deficient, however, in that it fails to recognize that the outcome of an election depends also on the quality of the campaign that candidates run. This quality depends on the amount of campaign contributions that they receive. To capture this feature of elections, we formulate the *pure media* model of Section IIB where we assume that the winning candidate is the one who attracts the most in campaign contributions. We also make a plausible assumption that a voter's contribution to a candidate is a function of the dollar welfare loss that his tax platform im-

*New York University. Our research has been funded by grants from the National Science Foundation and the Office of Naval Research. We should like to thank Sam Peltzman and Robert Willig for helpful comments. A more technical version of this paper is available from the authors.

[1] An earlier attempt at a similar problem was made by Schotter in the context of economic contraction.

[2] See William Baumol, Elizabeth Bailey, and Robert Willig for a complete discussion of sustainability and of conditions under which Ramsey-optimal prices are sustainable in a natural monopoly market.

poses on the voter in comparison to the loss inflicted by the other candidates. Hence, in the pure media model, politically sustainable tax vectors reflect the *intensity* of voter preferences over the feasible tax vectors. Surprisingly, we find that, in a two-candidate race, *only* the Ramsey-optimal taxes are politically sustainable. Unfortunately this result is quite sensitive to the number of competing candidates. For example, in a three-candidate race, there is no politically sustainable tax vector if candidates are allowed to choose identical platforms.

In Section IIC, we briefly discuss a model which incorporates the features of the two previous models. This model, called the *mixed-media* voting model, defines the outcomes of elections as a function of *both* the tax positions that candidates take as well as the amount of money they raise as a result of their platforms. Here, in a two-tax two-candidate race, if the median tax vector t^m is also (coincidentally) the Ramsey tax vector, then it is an equilibrium tax vector; otherwise, as is most likely, a politically sustainable tax vector, if it exists, is intermediate between t^m and t^{**}.

I. The Essential Model

In this section we provide a rudimentary review of optimal commodity taxation and develop those concepts which we shall utilize in the study of the voting models in the following sections.

Let us consider an economy which produces n final goods, $x_1, \ldots, x_n$, under conditions of constant returns to scale using labor l as the sole input. In such an economy, marginal cost pricing does not generate an overall budgetary surplus, hence governmental revenue requirements, if they exist, will necessitate an imposition of head taxes and distortionary taxes. In what follows we shall assume that income is untaxed leaving excise and head taxes as the only source of revenue.

Let us assume that consumers, indexed by $\alpha \in \Lambda$ where Λ is the index set of consumer-voters, have heterogeneous preferences, endowments of leisure time, and that they may differ in their productive efficiency. Let $y(\alpha) = p_0^\alpha T^\alpha + g$ be the full income of type α consumer, where T^α is his endowment of leisure time, p_0^α is his wage rate, and g is a lump sum transfer. Further if we let $p(\alpha) = (p_0^\alpha, p_1, \ldots, p_n)$ be the price vector facing that consumer, then $\mu^\alpha(p^0(\alpha)|p(\alpha), y(\alpha))$ is the Hurwicz-Uzawa income compensation function which gives the amount of income that a consumer of type α needs at some base prices, $p^0(\alpha)$, to make him as well off as he is at actual prices $p(\alpha)$ and full income $y(\alpha)$.[3]

The income-compensation function can be used to evaluate the welfare cost of a given tax and transfer program, and to define the maximum any voter would be willing to contribute to his favorite candidate. For instance, if we let $g=0$, c be the vector of producer prices, w^α be α's wage rate, and p be the vector of consumer prices, i.e., $p_i = c_i + t_i$, $i = 1, \ldots, n$, then μ^α gives the amount of income that α requires at some base price to make him as well off as he is after an imposition of a commodity tax vector t. Clearly, the higher is the value of μ^α, the less onerous is a given tax vector. Now let politician j, $j=1,2$, propose a tax vector t^j, (yielding consumer prices $p^j = c + t^j$). Then

$$(1)\quad z^\alpha(t^1, t^2; p^0(\alpha)) = \mu^\alpha(p^0(\alpha)|p^1, \cdot) - \mu^\alpha(p^0(\alpha)|p^2, \cdot)$$

gives the *maximum* amount that α would be willing to pay to ensure that the society selects t^1 rather than t^2. (Of course, if $z^\alpha < 0$ then α prefers t^2 to t^1 and would be willing to pay $|z^\alpha|$ to ensure the selection of t^2.)

We can state the social planner's problem as

$$(2)\quad \operatorname*{Max}_{t} \sum_{\alpha \in \Lambda} \mu^\alpha(p^0(\alpha)|c+t, w^\alpha, y(\alpha))$$

subject to the requirement that

$$(3)\qquad \phi(t) \equiv \sum_{i=1}^{n} t_i X_i(\cdot) \geqslant \bar{R}$$

where X_i is the aggregate market demand

[3] Willig suggested the use of the income compensation function to evaluate the efficiency costs of taxes.

for commodity i. The solution to the above program is the Ramsey-optimal tax vector t^*.

We now have all the necessary ingredients needed to study the various voting models defined above.

II. Politically Sustainable Tax Vectors

In this section we investigate whether politically sustainable tax vectors exist and when they do whether Ramsey-optimal tax vectors are politically sustainable. We show that the answers to those questions depend significantly on the nature of the voting model, on the number of candidates who are allowed to compete in the elections, and, of course, on the sources of heterogeneity among consumer-voters. We begin our discussion with a paradigmatic model of democracy wherein the winning politician is the one who attracts the required majority of voters.

A. *The Pure Democracy Voting Model*

Let us assume that two politicians are vying for voters in a general election. Each politician announces a feasible tax vector, that is, satisfying equation (3). Each voter casts his vote for that politician whose announced tax vector he most prefers. Hence, only the voter's ordinal preferences are expressed by his voting decision. It is intuitively clear that in this spatial-type model the politically sustainable (or equilibrium) tax vector, if it exists, will be determined by the preferences of the median voter, if a 50 percent majority is needed to win the election. The voting literature suggests that in spatial-type models, equilibria rarely exists. *Per force*, the same situation obtains in the pure democracy voting model.

It is not surprising, therefore, that stringent conditions are required for the Ramsey tax vector t^* to be identical with the equilibrium tax vector of the political process as defined by the median tax vector. The two instances in which this equivalence holds are given in

PROPOSITION 1: *Let $\Lambda \subset R^1$, then in the pure democracy model $t(\alpha^m)=t^{**}=t^*$ (where α^m is the median voter on R^1) if* (a) *all consumers have identical homothetic preferences and they differ only with respect to unearned income; or* (b) *consumers receive different wages and the individual utility function can be represented by $u(x_1, x_2, \ldots, x_n, l) = \phi(x) + al$ where ϕ is a homothetic function and l is labor supply.* (A proof of this and the following propositions is in our earlier paper.)

Condition (a) follows from the fact that homotheticity implies that budget shares $k_i = p_i x_i / y$, $i = 0, \ldots, n$, are independent of income. Consequently, for each consumer the same tax vector minimizes the burden of taxation as measured by the income compensation function. Condition (b) follows in fact from (a). Given the posited representation of the utility function, budget shares do not depend on the full income. Furthermore, labor supply can be viewed as being effectively fixed for each worker if workers are to supply any labor at all. Hence, cases (a) and (b) are essentially equivalent.

From this discussion we conclude that unless very restrictive preference structures and sources of heterogeneity among voters are imposed upon the pure democracy model, Ramsey-optimal taxes are not politically sustainable. This is not an entirely unexpected result in view of the fact that in this type of spatial voting model, equilibria rarely exist and the model is not sensitive to the cardinal preference intensities of the voters. The next section presents a model in which these preference intensities influence the election's outcomes.

B. *The Pure Media Voting Model*

In this model we posit that the winning candidate is the one who runs the best campaign. The quality of the campaign is a function of the money a candidate raises. Thus in the pure media voting model, the media alone determines the outcomes of political campaigns and the only function of political (tax) platforms is to raise money with which to wage a campaign and sway the voters. As in the previous section, we assume that each politician strives to maximize his probability of being elected. Hence each politician will propose that tax vector which, given the tax vector(s) proposed by his opponent(s), maximizes his campaign contributions.

Regarding contributions, we take it that in a two-candidate race, a candidate receives in contributions the amount equal to the maximum amount that voters would be willing to pay to *avoid* the election of the opposing candidate.[4] Thus, from equation (1), the jth candidate receives $Z^j = \Sigma_{\alpha \in C^j} |z^\alpha(t^1, t^2)|$ where C^j is the set of consumers who prefer the tax vector proposed by the jth candidate. With this assumption it can be shown

PROPOSITION 2: *In the pure media voting game with two candidates, the strategy of proposing the Ramsey tax vector t^* is a dominant strategy for each candidate. Furthermore, t^* is a unique equilibrium for that game.*

The intuition behind this result is that the Ramsey-optimal tax vector minimizes the social welfare loss as defined in equation (2) and, therefore, maximizes the sum z^j of campaign contributions made to the candidate who announces it for any tax vector proposed by the opposing candidate. Hence, if candidate 1 announces t^*, the second candidate's best response is to also announce t^* and thus guarantee himself a 50 percent chance of winning the election since no other t can guarantee him that much. If the incumbent politician is allowed to announce the tax vector first, and if the challenger is not allowed to replicate the incumbent's tax vector, then the incumbent can always ensure his reelection by selecting the Ramsey-optimal tax vector: the Ramsey-optimal tax vector is sustainable.

One important implication of the pure media voting model is

PROPOSITION 3: *If candidates are allowed to propose identical tax vectors then in equilibrium no candidate receives any contributions.*

If the challenger cannot imitate the incumbent's tax vector, the incumbent receives positive, albeit arbitrarily small, total contributions. Hence, in a pure media voting model, competition for voters drives campaign contributions towards zero.

Proposition 3 follows from the fact that when both politicians are proposing t^* (or any pair of identical tax vectors), consumer-voters are indifferent between them and hence find it unnecessary to contribute to either one. Furthermore, the candidate who deviates from that equilibrium can expect in contributions an amount that is smaller than the countervailing flow of contributions to the other candidate that is induced by his deviation.

It is instructive to explore whether the results reported in this section generalize to a $k>2$ candidate race. In general, we find that the incumbent can still assure reelection by proposing the Ramsey tax vector if challengers cannot replicate his tax program. However, with replication, the Ramsey tax vector is not an equilibrium tax vector. Thus,

PROPOSITION 4: *If individual preferences are strictly quasi concave, then in a k-candidate race, under appropriate continuity conditions, the Ramsey-optimal tax vector is not a voting equilibrium tax vector.*

This negative conclusion follows from the fact that if k-1 politicians announce the Ramsey tax vector t^*, then the remaining politician by announcing a tax vector $t^* + \varepsilon$, with ε arbitrarily small, can raise in contributions an amount smaller but nonetheless almost equal to the sum of contributions made to the remaining k-1 candidates. However, the k-1 candidates must split those contributions k-1 ways. Consequently the contribution of the deviant politician must exceed the per candidate campaign contributions of the remaining candidates and hence, by assumption he must win.

This result rests on the assumption that candidates can imitate each other's platform. If they cannot, we have

[4] This assumption rules out any type of free-rider problems that may arise on the voter side of the model. As we will see, however, *at the equilibrium* all such free-rider problems disappear. For another method of dealing with the free-rider problem, see Ordover and Willig.

PROPOSITION 5: *In a k-candidate race, the Ramsey tax vector is sustainable, that is, a candidate who announces it wins the election, if the other candidates cannot replicate his platform.*

Now that we have illustrated the pure democracy and pure media models let us investigate the mixed-media model, which is a hybrid of the two.

C. *The Mixed-Media Model*

Space does not permit a full description of the mixed-media model. Briefly, it is a model in which both the cardinal (intensity of) preferences and the ordinal preferences of the voters help to influence the outcome of the election. In the model, politicians first announce tax platforms. Upon these platforms the voters decide who they prefer (in terms of their ordinal preference) and also how much they want to contribute to the candidate they most prefer as defined by equation (1) (for a two-candidate race). The candidates then use these contributions to run their campaign with the assumption that the candidate who raises the most in campaign contributions will be able to steal some of the voters who ordinally prefer the other candidate by engaging in a media campaign. This possibility is summarized by a media technology as represented by a "stealing function" (a function of the excess of contributions of the candidate with the largest campaign chest over his competitor). Hence the model is a hybrid of the pure democracy and the media models.

The results we derive are of interest because they alter the median voter result typically encountered in the voting literature. To summarize these results consider the following two propositions.

PROPOSITION 6: *In a two-candidate n-tax race, the median tax vector is an equilibrium tax vector if it is also the Ramsey tax vector.*

Most often, however, the Ramsey tax vector is not identical to the median tax vector. When this is the case, we are able to establish the following result:

PROPOSITION 7: *In a two-candidate two-tax race, where t^* is the Ramsey tax vector and t^m is the median tax vector, with $t^* \neq t^m$, and in which candidates compete by setting t_1 only (t_2 is then defined from the revenue constraint* (3)), *then*; 1) *if an equilibrium exists it must be an equilibrium in which both candidates set the same tax vector t^{**}, and* 2) *if $t_1^* \lesseqgtr t_1^m$, then $t_1^* \lesseqgtr t_1^{**} \lesseqgtr t_1^m$. In other words, the equilibrium tax vector of the mixed-media game lies betwen the Ramsey and median tax vector.*

III. Conclusions

To conclude our discussion we can state that, unlike the system of economic competition, political competition is unlikely to lead to first (or even second) best societal outcomes. The more that cardinal intensities can be represented in the voting game described by the political process, however, the more likely we are (as in the pure media game) to have the politically sustainable outcomes of the process be economically efficient.

REFERENCES

Anthony B. Atkinson and Joseph E. Stiglitz, *Lectures on Public Economics*, New York: McGraw-Hill 1980.

W. J. Baumol, E. E. Bailey, and R. D. Willig, "Weak Invisible Hand Theorems on the Sustainability of Multiproduct Monopoly," *Amer. Econ. Rev.*, June 1977, *67*, 350–65.

J. A. Ordover and A. Schotter, "On the Political Sustainability of Taxes," mimeo., Sept. 1980.

______ **and R. Willig,** "Money is the Message: Towards Political Economy of Campaign Contributions," mimeo., Aug. 1980.

A. Schotter, "Economically Efficient and Politically Sustainable Economic Contractions" in R. Henn and O. Moeschlin, eds., *Mathematical Economics and Game Theory*, Berlin 1977.

U.S. Incomes Policies in the 1970's—Underlying Assumptions, Objectives, Results

By D. QUINN MILLS*

The 1970's opened with a program of wage restraint in the construction industry. The effort was tripartite, and made minimal use of the regulatory authority of the government. Thereafter, a general wage-price freeze opened a program of primarily governmental restraint.

When the broader controls efforts were established under an agency established in the Executive Office of the President, the construction effort chose to remain in the Labor Department largely independent of the larger controls program. As a result, the official histories of the Pay Board and Price Commission (1971–72) exclude that of the construction program, whose history is instead reported by the Labor Department and is far less well known to economists and congressional officials today. A short period of tripartitism in the Pay Board gave way to a primarily government effort. Thus, the 1970's opened with a conflict between the concept of an industry-specific and tripartite program on the one hand, and the concept of a generalized, regulatory effort of the government on the other.

An incomes policy of the type embedded in the Republican administration's Pay Board and Price Commission was swept aside in 1974, only to reappear in 1978 under a Democratic administration. In the interim, a less-formal policy without published regulations or threatened sanctions had been pursued. In sum, beginning in March 1971, the federal government has utilized some form of incomes policy virtually continually.

Still, the decade was marked most clearly by reliance on a certain type of policy—one which was essentially a unilateral effort of the government, not conducted in cooperation with business and labor, which took the form of a regulatory program, and which treated the rate of increase in overall compensation as the key policy variable to which the program was addressed. What are the assumptions that motivate this type of program? How have objectives been set, and what have been the results?

This paper cannot cover exhaustively the assumptions, objectives, and results of the various programs which have existed in the 1970's. Instead, I shall concentrate on the Carter Administration's program of wage and price guideposts for October 1978 to October 1979. This program drew for its basic features and its top personnel on the Pay Board and Price Commission of 1971–72. But its authors in the Carter Administration attempted to develop a program which was corrected for the faults which they had observed in the earlier effort.

I. Assumptions

A. *How Wages and Prices Interact*

According to the Council on Wage and Price Stability (CWPS) (which administers the Carter Administration incomes policy), the 1978–79 program involved a clear nexus between the pay and price standards: specifically, the first-year price standard was derived from the pay standard, assuming a constant percentage markup of price over unit labor costs. The pay standard was 7 percent. CWPS added one-half of a percentage point because of relatively large increases in employment taxes and subtracted one and three-quarters percentage points for trend productivity growth. The 5–3/4 percent aggregate price standard was one-half of a percentage point less than price increases during the 1976–77 base period. Therefore CWPS set a company-specific

*Graduate School of Business Administration, Harvard University.

"price deceleration" standard which called for limiting price increases to one-half of a percent less than the base-period rate of change. (CWPS Memorandum to the Price Advisory Committee, February 27, 1980.)

A similar formulation had been used by the Pay Board in 1971 in establishing its general standard for pay increases. The target for price increases was 2.5 percent; anticipated productivity growth was 3.0 percent. The sum of price increases and productivity growth yielded the general pay standard of 5.5 percent.

It is the high proportion of employee compensation as part of total business costs which motivates this type of policy formulation. As Daniel J. B. Mitchell, former chief economist of the Pay Board, said in testimony before the House Budget Committee, June 26, 1979, "Structural factors ensure that wages will be the central element of a guidelines...program, even though the ultimate target is price inflation."

But this emphasis on wage restraint in 1978–79 now appears somewhat anamolous, when there has been general recognition that energy, food, and interest charges have each been major contributors to inflation, but largely outside the purview of the guideposts themselves.

B. *How Wages are Determined*

The current guideposts program of 1978–79 attempted to control the average rate of increase in compensation. But in the United States, institutions of wage setting are so decentralized that average compensation cannot be a policy variable. To echo George Meany's oft-repeated lesson to President Johnson, the United States is not Sweden. There is not in this country a national pay bargain which the government can affect by its regulatory or other efforts.

In consequence, in the United States an incomes policy is, in practice, by necessity a program of intervention at the micro level of the economy. Standards must be set not for broad, overall decisions about pay, but for many decentralized actions. The method by which macro-level objectives are translated into micro-level standards is central to the design and administration of any incomes policy in the United States.

What has been the contribution of contemporary wage theory to developing a methodology to effect this translation? At the micro level, human capital theory has spawned many studies of wages and income levels in which the primary explanatory variables are demographic characteristics. At the core of human capital theory is the observation that the compensation of individuals typically rises with age and longevity. In consequence, theorists have jumped to the conclusion that this primarily reflects increasing investment in the skills of individuals. But observations of racial and sexual discrimination in pay, and of the internal pay practices of companies (which show that performance measures are not directly correlated with seniority or pay), contradicted this.

Empirical observations suggest instead a micro-economic theory of wages which includes institutional, attitudinal, and behavioral factors, as well as labor and product market factors. Some theories which stress these factors also stress wage relationship and comparisons (or contours). These theories are sometimes interpreted as theories of the general level of wages, but are not. Instead, they are theories of wage determination at the micro level. As such, they are useful in an incomes policy's application. Human capital theory is not.

At the macro level, we have little more as a theory of wages than the virtual tautology which states the relationship between changes in the general pay level, the general price level, and the general level of labor productivity. It is not difficult to base an educational program on this tautology, as did the Council of Economic Advisors in 1962–65, but it is hazardous to base an incomes policy on it, as has been attempted repeatedly in the 1970's.

These attempts have largely failed. In essence, the government first attempted in 1971–72 to apply a single percentage standard to all compensation increases. A question of feasibility in our decentralized system quickly arose. The effort to be rigid in support of the general standard gave way almost immediately to a much more flexible application of the standard. But flexibility can be introduced in either of two ways. One way is to permit specific exceptions to the general standard. A second is to calculate the costs of compensation increases differently. An example will make clear the difference. In 1978, the Teamsters and Trucking Management, Inc. negotiated a Master Freight Agreement (covering some 350,000 employees) to apply in 1978–81. The Agreement was estimated by the employer and the union to cost about 32 percent over the three years. The guideposts allowed an increase of 22.5 percent. How was the government to respond? One way would be to permit the Agreement as an exception to the general standard, but recognize it as a 32 percent agreement. The other way was to omit certain items in the Agreement, and use conventions which understate the cost of other items, so that there is introduced a "guideposts arithmetic" which contrasts with the "actual arithmetic" (these terms, or their equivalents, are in common usage now in the administration). The Carter Administration opted for the second course. Therefore, in the spring of 1979, CWPS announced that it had costed the new National Master Freight Agreement at 22.5 percent, and found it in compliance with the guideposts. This result was obtained primarily by costing the negotiated cost of living allowance (*COLA*) as if inflation would be 6 percent per year for three years, despite rates of inflation which were considerably higher, and by excluding from the calculation part of the initial wage increase due under the contract and all of the final increase due as the agreement expired.

Thus the question of flexibility in a program utilizing a general standard for pay increases has apparently been settled. There is to be considerable flexibility. The issue now is how that flexibility is to be introduced. Some observers prefer the use of actual arithmetic and case-by-case rulings on requests for exceptions. Others prefer the use of guideposts arithmetic, so that the apparent applicability of the general standard to all situations can be maintained (possibly as a device to enhance the psychological impact of the program and to enhance its apparent evenhandedness in application).

There are certain consequences of guidepost arithmetic which go beyond introducing flexibility into the program, however. Guidepost arithmetic discriminates strongly against certain employee groups. Many employees lack the protection of *COLA* clauses, or are in small units whose employers or unions, if any, are not sophisticated enough to apply guidepost arithmetic to their own advantage. These groups end up being judged before the standards by ordinary arithmetic, and are sometimes forced to reduce pay increases to conform to the guideposts. However, these employees may have actually received far less in compensation increases than others who benefited from the guidepost arithmetic.

C. *The Components of Compensation*

Current guidepost policy involves a certain paradox concerning the components of compensation: salaries and benefits. In order to measure pay increases against the general standard, current policy treats all benefits and wages as essentially alike, reducing each to employer cost and measuring the cost of the compensation package as a whole. Yet, in order to introduce flexibility into the program through its costing procedures, the administrators treat many benefits and pay practices differently in the cost procedure. Since some benefits and pay practices are costed differently, some are favored and others are not. In the current program, especially favorable cost treatment has been accorded to *COLA*s. As we shall see, this was not intended, but was a result of the contradiction between intending to focus only on total compensation cost, while simultaneously treating items of compensation differently. It is a problem inherent in any program attempting to regulate primarily the total cost to the employer of a compensation package.

The results of this problem are borne disproportionately in the economy. For industries such as construction, which do not use *COLA*s, there is no alternative but to accept lower pay increases than in industries which use *COLA*s, or violate the guideposts.

Nor is the problem confined to *COLA*s. Pension programs differ markedly in the benefits that can be provided for any given amount of cost. The difference in benefits arises from variation in the demographic characteristics of the group covered by a pension plan. A guidepost program which uses a cost-only standard for pensions necessarily condemns some employees to lower benefit levels than others.

Many people see results such as these as inequitable. As the apparent inequities become well known, the guidepost program begins to lose its creditility and its effectiveness.

II. Objectives

A. *Concerning the Economic Environment*

During World War II direct wage and price controls were justified as necessary to keep excessive aggregate demand from generating rapid inflation and disorderly conditions in labor markets. Presumably in a slack economy market forces would restrict inflation. But in the 1970's, direct restraint has become a man for all seasons. In the words of the CWPS *Inflation Update* (November 23, 1979), "pay and price standards can be expected to be most effective in moderating wage and price increases in an economy characterized by slack rather than tight market conditions."

The fact is that direct controls, or guideposts, are now thought of as a generalized element of economic policy, suitable to slack as well as tight market conditions. This change in attitude is marked among government economists, in the administration and in Congress. It increases dramatically the likelihood of continuing programs of direct restraint.

B. *Concerning the Function of Incomes Policy*

Paradoxically, again, the acceptance of direct restraint as a generalized device has been accompanied by a reduction in the scope of its function. During World War II, fiscal monetary and incomes (then "stabilization") policies were each viewed as tools of multiple effect, to be employed together in pursuit of an overall objective which had several elements. Today, incomes policy is thought of as a tool of specialized effect. Its purpose is to limit compensation increases. Its demonstrated capability (as we shall see later) is primarily against the pay increases of nonunionized employees. Its potential, or actual, effect on the income distribution is largely ignored (although its possible effect on income shares between capital and labor is closely monitored). Its inevitable effects on collective bargaining, industrial relations, productivity, industrial revitalization, and the degree of union organization seem accidental to the administrators rather than intended.

C. *Concerning the Impact of Prices on Wages*

Although as noted above the rationale for the guideposts runs from controlling wages to controlling prices, the actual objectives of the program have been to prevent so-called exogenous price increases (in food, interest, and energy) from accelerating the rate of compensation increases. The CWPS noted in its November 1979 *Inflation Alert* that, "The surges in fuel, housing and food-price inflation... have not yet become built into the industrial wage/price structure." Apparently, the purpose of the guideposts was to prevent the price surge from being reflected in wages.

III. Results

The CWPS cites evidence that the 1978–79 pay standards have had some success in lessening the rate of inflation overall. In its most recent pronouncement, CWPS estimates that the Council's guideposts reduced the rate of inflation in wages from October 1978 to March 1980 by 1.8 to 2.0 percentage points and reduced the rate of price inflation by 1.1 to 1.5 percentage points in the same period (*Interim Report*, May 6, 1980).

Leaving aside the question of the econometric validity of the CWPS estimates, the primary reason for CWPS's alleged success is the limitation the guideposts apparently have placed on *nonunion* wage increases. That there was such a limitation seems reasonably clear from the statistical evidence.

This was, however, a phyrric victory for the program. As nonunion pay increases were squeezed in the first program year (October 1978 to October 1979), employers became concerned about possible unionization efforts among nonunion production, clerical and professional workers, and about compression of supervisors' salaries as compared to unionized production workers' earnings (points made by Albert Rees in congressional testimony on March 25, 1980). In consequence employers sought substantial liberalization in terms of the pay guideposts for the 1979–80 period. In part because employers joined unions in opposition to the continuation of the program into 1979–80, the administration created a tripartite Pay Advisory Committee which recommended and obtained substantial modification of the pay guideposts over the opposition of the CWPS staff. The changes then introduced into the incomes policy were described by Chairman John Dunlop of the Pay Advisory Committee in Hearings before the House Committee on Banking, Finance and Urban Affairs, March 19, 26, and May 6, 1980, pp. 293–313.

How the Council got itself into this position is instructive, because it set out to do the opposite. In its paper of November 23, 1979, CWPS indicated its view that "the underlying rate of inflation reflects the extent to which the economy is effectively indexed." In order to limit inflation, by implication, CWPS sought to reduce indexation. One of the major elements of indexation in our economy is the existence of COLA clauses in many collective bargaining agreements. Yet the regulations adopted by CWPS gave strongly favorable treatment to *COLA*s by costing the escalators as if the rate of consume price inflation were only 6 percent in 1979, instead of the 13 percent which actually occurred. The sources of this self-defeating regulatory treatment are several and too complicated to be explored here.

IV. Conclusion

In the late 1970's, incomes policies have become increasingly a narrowly focused adjunct tool of economic policy. In a sense, CWPS has become a clone of the Council of Economic Advisors concerned with macroeconomic policy, but not with industrial relations, the income distribution, productivity, or industrial revitalization policy.

CWPS has mirrored the CEA in developing its capacity for macro-economic analysis, including the diagnosis of economic problems and the establishment of short-term economic objectives. It has been very weak in terms of the design and implementation of its policy. Because in the implementation of overall economic policy CWPS must interface with decentralized processes and actors in the operation of individual product and labor markets, its macro focus is a serious handicap. It is hard to fault CWPS's definition of its mission in overall economic policy as set forth in its November 1979 paper. But it is all too easy to fault its efforts to carry out its mission.

This is unfortunate since an incomes policy run with a narrow macro-economic focus cannot build the public support, or the constituency, necessary to survive for any significant length of time. Instead, incomes policy becomes subject to sudden and confusing shifts in policy and staff, and much of its effectiveness is lost.

In part, CWPS's weakness reflects a limitation in the economics profession. An incomes policy is an exercise in applied economics at a time when applied economics is attracting little attention in the training of economists. An incomes policy is an exercise in microeconomics, at a time when macroeconomics gets most attention. An incomes policy is an exercise in the economics of particular industries, occupations and geographic areas, at a time when industrial, occupational and geographic studies are not widely pursued. Therefore, the development of macroeconomics in recent decades has taught us the importance of utilizing income policies, while the relative atrophy of applied microeconomics in labor and product markets has deprived us of the capability to conduct income policies well.

Government Intervention in the Inflation Process: The Econometrics of "Self-Inflicted Wounds"

By JON FRYE AND ROBERT J. GORDON*

The high variance and continued acceleration of inflation during the 1970's pose new challenges to the time-series econometrician. The theme of this paper is that inflation in the past decade has depended not only on the level of aggregate demand and the role of inertia—the two explanatory variables stressed in the conventional Phillips curve framework—but also on a number of different supply shocks. Two of these, the increase in the relative price of oil and decline in the rate of productivity growth, have been outside of the direct control of the government, particularly in the short run. But both the variance and acceleration of inflation have been aggravated by three measures within the purview of government policymakers: increases in the effective Social Security tax, increases in the minimum wage rate, and episodes of direct government intervention in the price-setting process. Because of their futility, these intervention episodes can be regarded as "self-inflicted wounds," like the tax and minimum wage changes that normally are described by this term.

I. Basic Specification of the Reduced-Form Inflation Equation

We begin from a pair of wage and price equations and combine them to obtain our basic reduced-form equation that is used for estimation below. The rate of wage change depends on the sum of lagged price change and the desired rate of real wage growth, on the level and rate of change of the output ratio, and on supply shifts that affect the wage-setting process. The rate of price change relative to the current rate of wage change depends on the change in "standard" productivity, the level and rate of change of the output ratio, and on supply shifts that affect the price-setting process. When these two equations are combined, we obtain

$$(1)\qquad p_t=\gamma_0 p_{t-1}+\gamma_0(\lambda_t-\sigma_t)+\gamma_1\hat{Q}_t+\gamma_2\hat{q}_t+\gamma_3 z_t+\varepsilon_t$$

where uppercase letters designate *logs* of levels of variables and lowercase letters designate their proportional rates of change. Equation (1) states that the inflation rate (p_t) depends on past inflation (p_{t-1}), the difference between the desired rate of real wage growth in the wage equation (λ_t) and the rate of standard productivity growth relevant for price-setting decisions (σ_t), the level of the output ratio ($\hat{Q}_t$), the rate of change of the output ratio ($\hat{q}_t$), a vector of supply shift variables (z_t), and an error term (ε_t).[1]

There is one rather subtle obstacle to the estimation of (1). We would expect the rate of inflation to respond positively to the speed of economic expansion, $\hat{q}_t$. But there are two reasons why p_t and $\hat{q}_t$ may have a negative correlation that results in a downward bias in the coefficient γ_2. One reason is measurement error; since nominal *GNP* and prices are measured independently, with real *GNP* as a residual, any error in the measurement of prices introduces an opposite movement in $\hat{q}_t$. Second, for any given growth rate of nominal *GNP*, a supply shock ($z_t>0$) raises p_t and reduces $\hat{q}_t$; any errors in measurement of the z_t variables may introduce a spurious negative correlation between p_t and

*Northwestern University. This research has been supported by the National Science Foundation. We are indebted to Elizabeth H. Johnson and Joan Robinson for their indispensable help in the preparation of this paper.

[1] A more detailed development of both this specification and of the subsequent empirical results is contained in our earlier paper.

$\hat{q}_t$. To avoid this problem we use the identity $p_t + \hat{q}_t = \hat{y}_t$, where the latter variable stands for the excess of nominal *GNP* growth over the growth in natural real *GNP* ($\hat{y}_t = y_t - q_t^*$). When this identity is substituted for $\hat{q}_t$ in (1), we can factor out p_t and obtain our final estimating equation:[2]

$$(2)\quad p_t = \frac{1}{1+\gamma_2}\Big[\gamma_0 p_{t-1} + \gamma_1 \hat{Q}_t + \gamma_2 \hat{y}_t + \gamma_0(\lambda_t - \sigma_t) + \gamma_3 z_t + \varepsilon_t\Big]$$

In long-run equilibrium, inflation (p_t, p_{t-1}) and adjusted nominal *GNP* growth ($\hat{y}_t$) are equal and all other variables are zero. This implies that the sum of coefficients on lagged inflation and adjusted nominal *GNP* growth must be unity to allow this steady-state equilibrium to be attained.

II. Two Unicausal Approaches

We first provide estimates of two simpler equations that stress single-cause explanations of inflation. In recent years considerable attention has been given to autoregressive integrated moving average (*ARIMA*) models which represent an extreme view that the inflation process is entirely dominated by inertia and is unaffected by changes in current exogenous variables (see Edgar Feige and Douglas Pearce). Another unicausal approach is a simple monetarist equation that makes the rate of change of prices depend only on a distributed lag of past changes in the money supply. While this framework is taken more seriously by journalists and laymen than academic economists, a "money only" explanation of inflation is implicit in some recent tests of the classical equilibrium approach to macroeconomics.[3]

We use the *ARIMA* and money-only equations to provide an alternative estimate of the effect of the Nixon price controls.[4] Columns (1) and (2) of Table 1 display the resulting coefficients on the dummy variables and the summary regression statistics. Both the *ARIMA* and money-only models fit the data for the 1954–80 period with similar standard errors of about one percentage point. The Nixon controls dummy variables are scaled to show the cumulative impact of the controls on the price *level* during the appropriate period, and thus their coefficients in both columns (1) and (2) indicate that the controls held down the price level by about 3 percent at the end of 1972, while their termination allowed the price level to bounce back to roughly its no-controls level.

An alternative method of assessing the impact of controls is to compute a postsample dynamic simulation of an equation estimated to the precontrols period and treat it as an estimate of inflation in the counterfactual state.[5] Lines 14a and 14b of Table 1 show the postsample simulation errors of an equation estimated for 1954:2 to 1971:2.

III. Specification and Results for the Basic Equation

The third column of Table 1 presents estimates of our basic equation as specified in equation (2) above and exhibits a standard error of 0.68, little more than half that of the unicausal models. A line-by-line discussion of our variables and results follows:

1) *Lagged Inflation*. The inertia in the inflation process is captured by a distributed lag on 24 past values of fixed-weight *GNP* deflator inflation. Because the explicitly temporary effects of the controls program should not enter this measure of inertia, the estimated controls effects are removed from the lagged dependent variable, requiring iterative estimation.

[2] Equation (2) contains productivity and supply shift terms but otherwise is identical to equation (6) in Gordon's 1980 paper.

[3] See especially the paper by Robert Barro and Mark Rush.

[4] Our use of dummy variables to assess an intervention in an *ARIMA* process follows the procedures suggested by G. E. P. Box and G. C. Tiao.

[5] For more on the methodology of estimating the impact of controls and other types of government intervention, see Gordon (1973), Alan Blinder, and Walter Oi.

TABLE 1—MEASURES OF THE IMPACT OF NIXON-ERA WAGE AND PRICE CONTROLS USING ALTERNATIVE MODELS OF THE INFLATION PROCESS FOR THE PERIOD 1954:2-1980:2[a]

	ARIMA	Money-Only	Comprehensive
	Model[b]	Model	Reduced Form
	(1)	(2)	(3)
1) Lagged Inflation[c]	–	–	0.90
			(15.4)
2) "On" Dummy[d] 1971:3-1972:4	−3.14	−3.31	−1.30
	(−2.99)	(−4.68)	(−2.65)
3) "Off" Dummy[d] 1974:2-1975:1	2.46	3.07	1.60
	(3.40)	(5.07)	(2.30)
4) Current and Lagged *M*-1B[e]	–	1.46	–
		(21.8)	
5) Output Ratio ($\hat{Q}_t$)	–	–	0.19
			(3.55)
6) Adjusted Nominal *GNP* Growth ($\hat{y}_t$)	–	–	0.14
			(4.56)
7) Food and Energy Prices	–	–	0.29
			(4.22)
8) Productivity Deviation[f]	–	–	−0.38
			(−5.47)
9) Effective Exchange Rate[f]	–	–	−0.09
			(−3.21)
10) Social Security Tax[f]	–	–	0.54
			(2.98)
11) Effective Minimum Wage Rate[f]	–	–	0.02
			(1.66)
12) Constant	0.16	−1.42	–
	(1.42)	(−5.43)	–
13) a. *S.E.E.*	1.15	1.05	0.68
b. *D.W.*	2.07	1.60	2.19
14) Cumulated Errors from Dynamic Simulation within Specified Intervals[g]			
a. "On" 1971:3-1972:4	−1.93	−3.46	−1.23
b. "Off" 1974:2-1975:1	5.28	4.09	3.08

Note: Numbers in parentheses are *t*-statistics.

[a] The dependent variable is 400 times the quarterly first difference of the *log* of the fixed weight *GNP* deflator.

[b] The coefficients in column (1) are estimated in a regression equation in which all variables are pre-filtered.

[c] The coefficient shown is the sum of 24 distributed lag coefficients constrained to lie along a fourth-degree polynomial with a zero end constraint.

[d] The dummy variables are constrained to add up to 4.0 (reflecting the conversion of quarterly changes of all variables to annual rates). Thus the "on" dummy is equal to 2/3 for the six quarters listed, and the "off" dummy is equal to 1.0 for the four quarters listed.

[e] The coefficient shown is the sum of 28 distributed lag coefficients constrained to lie on a fifth-degree polynomial with zero end constraint.

[f] The coefficient shown is the sum of a set of unconstrained coefficients on current and lagged values, with four lags included on lines 8, 10, and 11, and two lags included on line 9.

[g] The equation represented by each column is reestimated for the period 1954:2-1971:2 and dynamically simulated beginning 1971:3. In column (3) estimation is subject to the constraint that the sum of coefficients on adjusted nominal *GNP* growth and lagged inflation equals 1.

2); 3) *Nixon Control Dummies*. The Nixon controls program is estimated to have held down prices 1.30 percent at the end of 1972, but this effect was more than cancelled by the rebound inflation of 1.60 percent. The estimate of each effect is about half of the corresponding estimate in the unicausal models. This is because the unicausal models must attribute the control period effects of all omitted variables to the control dummies. But inflation was low in 1971:3–1972:4 in part because of the productivity gains of this period, and inflation was high in 1974 in part because of a productivity reversal, food and energy price shocks, and the depreciation of the exchange rate of the *U.S.* dollar.

5) *Output Ratio*. This variable is the *log* of the ratio of real output to the natural rate of output, i.e., $\hat{Q}_t = Q_t - Q_t^*$. The Q^* variable used to obtain the output ratio and, in rate of change form, to adjust nominal *GNP* growth is from Jeffrey Perloff and Michael Wachter. This traditional Phillips curve variable is highly significant; its coefficient of 0.19 indicates that a one percentage point excess of actual real *GNP* above natural real *GNP* causes an acceleration of inflation of 0.19 percentage points at an annual rate per quarter.

6) *Adjusted Nominal GNP Growth*. The nominal *GNP* growth variable is defined net of natural real *GNP* growth. A slowdown in the trend growth rate of productivity will reduce natural real *GNP* growth and raise $\hat{y}_t$, so that this variable represents the combined effects of demand stimulation and trend productivity growth. The implied parameter estimates are $\gamma_1 = 0.22$ and $\gamma_2 = 0.16$.

7) *Relative Prices of Food and Energy*. The contribution to inflation of changes in the relative prices of food and energy is measured by the difference between the rate of change of the private business deflator and that of an alternative deflator that attempts to "strip out" the impact of the changing relative prices of food and energy. While this variable makes a significant contribution to the fit of the equation, its coefficient indicates that only a fraction of the relative prices changes was incorporated into a permanent acceleration of inflation.

8) *Productivity Deviation*. The variable standing for $\lambda_t - \sigma_t$ is the deviation of actual productivity growth from the productivity trend, estimated to be a constant for 1954–69 and a declining time trend during 1970–80. Its coefficient indicates that the productivity variable used in price setting (σ_t) is an average based 38 percent on actual productivity changes and 62 percent on the productivity trend.

9) *Effective Exchange Rate*. The depreciation of the dollar during the 1970's has been excluded or statistically insignificant in previous studies. This previous insignificance stems from the impact of the Nixon controls in delaying the adjustment of *U.S.* domestic prices to the dollar depreciation that occurred in two stages between 1971 and 1973. We have created a new variable which is equal to the actual change in the effective exchange rate of the dollar starting in 1974:3, which is set equal to zero before 1974, and which in 1974:1 and 1974:2 equals the cumulative depreciation that occurred between 1971:3 and 1974:2. Its coefficient indicates that a 10 percent dollar depreciation raises the inflation rate by 0.9 percentage points in the first three quarters.

10) *Social Security Tax*. The coefficient of 0.54 indicates that half of all changes in the effective tax rate,[6] which includes both employer and employee shares, is shifted forward into prices.

11) *Effective Minimum Wage Rate*. This variable is defined as the ratio of the statutory minimum wage to average hourly earnings in the nonfarm private economy. Its coefficient of 0.02 means that the cumulative 8 percent increase in the effective minimum wage rate during the four quarters in 1978 accounted for an acceleration of inflation of about 0.16 percentage points.

An alternative assessment of the effect of controls is provided by the dynamic simulation beginning 1971:3 of our basic equation fit to data through 1971:2. The "on" effect estimated by dynamic simulation and reported on line 14a approximates the dummy variable estimate, but the estimated "off" effect is much higher, because the pre-1971:3 equation does not contain the effective foreign exchange rate. The post-sample simulation incorrectly attributes the inflationary impact of the depreciation of the dollar to the removal of the controls program. To correct for this, we have run two in-sample dynamic simulations of the 1954–80 equation, one of which sets the change in the effective exchange rate to zero. The difference between the two simu-

[6] The variable is calculated as the percentage change in $(1/1-\tau)$, where τ is the ratio of total federal and state Social Security contributions to total wage and salary income in the national income accounts.

lations yields the estimate that 1.50 percentage points of the high inflation of the off period was contributed by the foreign exchange variable. A more credible estimate of the impact of the termination of controls is therefore 3.08 − 1.50 = 1.58, which approximates the dummy variable estimate of 1.60.

IV. Sensitivity and Extensions of the Basic Equation

Another episode of government intervention occurred during the Kennedy and Johnson Administration, when there were quasi-voluntary guidelines established for wage increases. These guidelines, first mentioned in the 1962 Economic Report of the President, are assumed to be in effect between 1963:1–1965:4. We enter a separate dummy variable for the three-year period beginning in 1966:1 to assess the possibility that part of the 1966–68 acceleration in the inflation rate was due to the end of the guidelines rather than a general state of excess demand in the economy. When these dummy variables are included in our basic equation, the resulting coefficients and *t* statistics are:

Guidelines dummy I (1963:1–1965:4)
0.01 (0.01)
Guidelines dummy II (1966:1–1968:4)
0.60 (0.61)

The verdict of these coefficients is that the guidelines program had no significant effect on inflation. The positive influence on inflation of demand growth in the 1964–65 period was offset not by the guidelines program, but by rapid productivity growth. An important implication of this result is that if the guidelines had a significant effect in holding down wage increases, then the program created a boom in the profit share.

The Carter pay standards may be similarly assessed. We introduce two dummy variables for the periods 1978:4–1979:4 and 1980:1–1980:2, respectively. The resulting coefficients and *t* statistics are:

Carter dummy I (1978:4–1979:4)
−0.67 (−1.08)
Carter dummy II (1980:1–1980:2)
0.05 (0.18)

Both variables are insignificantly different from zero, suggesting that there was nothing unusual about the inflation experience between late 1978 and mid-1980, and that the other variables in the equation are capable of tracking the data.

An alternative method of assessing the Nixon controls introduced by Alan Blinder and William Newton estimates an equation which does not use dummy variables. Rather, a new variable that represents the on effect is equal to the fraction of the *CPI* subject to price controls in each month, based on government records. We substitute the Blinder-Newton on variables and current and four lagged values of the off variable for our control dummies, and, following Blinder and Newton, assess the controls effects by two dynamic simulations, one of which has the controls variables set to zero. The implied controls effect (column (a)) may be compared to our own from Table 1 (column (b)).

	(a)	(b)
Standard Error	0.68	0.68
Maximum Restraint of Inflation	−1.48%	−1.30%
Postcontrols Rebound	+1.35%	+1.60%

The Blinder-Newton technique—despite the extra research required for construction of the new variable and its lack of applicability to other episodes of government intervention—provides neither a better fit nor an evaluation of the Nixon controls that differs from our simple dummy variable approach.

V. Conclusions

An adequate explanation of inflation in the 1970's requires a model that includes inertia in the adjustment of prices and the effects of aggregate demand, external supply shocks, and government intervention. Our basic reduced-form inflation equation relies on the contribution of two variables for its aggregate demand effect, the level of the output ratio, and the change in nominal

GNP adjusted for changes in natural real *GNP*. External supply shocks include changes in the relative prices of food and energy, the influence of changes in the effective exchange rate of the dollar, and deviations of productivity from trend. Three forms of government intervention influence inflation, the Nixon-era controls, as well as changes in the effective Social Security tax rate and effective minimum wage.

Three different methods are used to assess the impact of the Nixon-era controls within the context of our basic reduced-form inflation equation. Postsample dynamic simulations tend to underpredict inflation in 1974 more than they overpredict inflation in 1972, partly because there was no role of the effective exchange rate before 1971. The inclusion of dummy variables for the imposition and removal of the controls has the advantage of using all of the information available in the full sample period. Dummy variables indicate that the Nixon controls held down the price level by about 1.3 percent between mid-1971 and late 1972, and then allowed a rebound of about 1.6 percent to occur in 1974 and early 1975. A third technique, introduced by Blinder, seems conceptually superior, but it does not alter the conclusions of the dummy variable technique.

Why was inflation so variable between 1971 and 1980? And why did inflation accelerate from 5 percent in early 1971 to 10 percent in early 1980? Our basic equation explains the high variance of inflation mainly as a result of swings in the effect of Nixon controls, the deviation of productivity from trend, the relative prices of food and energy, and the effective exchange rate, with an additional minor contribution made by the aggregate demand variables and by Social Security tax changes. The overall acceleration of inflation during the past decade is explained by the adverse contribution of most of the variables.

While the inflation equation developed in this paper identifies the main factors that explain the recent behavior of inflation in the United States, additional research is required before this framework can be used to assess the consequences of alternative aggregate demand policies. A restrictive demand policy, for instance, would alter the inflation rate not only through the nominal *GNP* growth and output ratio variables, but also through the effect of demand policy on the behavior of productivity and the exchange rate, requiring that auxiliary equations be estimated to capture these indirect channels of influence.

REFERENCES

R. Barro and M. Rush, "Unanticipated Money and Economic Activity," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago 1980.

Alan Blinder, *Economic Policy and the Great Stagflation*, New York 1979.

______ **and W. Newton,** "The 1971-74 Controls Program and the Price Level: An Econometric Post-Mortem," Nat. Bur. Econ. Res. work. paper no. 279, 1978.

G. E. P. Box and G. C. Tiao, "Intervention Analysis with Applications to Economic and Environmental Problems," *J. Amer. Statist. Assn.*, Mar. 1975, *70*, 70–79.

E. Feige and D. Pearce, "Inflation and Incomes Policy: An Application of Time Series Models," in Karl Brunner and Alan H. Meltzer, eds., *Economics of Wage and Price Controls*, Vol. 2, Carnegie-Rochester Conferences on Public Policy, *J. Monet. Econ.*, Suppl. 1976, 273–302.

J. Frye and R. J. Gordon "The Variance and Acceleration of Inflation in the 1970s: Alternative Explanatory Models and Methods," Nat. Bur. Econ. Res. work. paper no. 551, 1980.

R. J. Gordon, "The Responses of Wages and Prices to the First Two Years of Controls," *Brookings Papers*, Washington 1973, *4*, 765–79.

______, "A Consistent Characterization of a Near-Century of Price Behavior," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 243–49.

W. Oi, "On Measuring the Impact of Wage-Price Controls: A Critical Appraisal," in Karl Brunner and Alan Meltzer, *Economics of Wage and Price Controls*, Vol. 2, Carnegie-Rochester Conferences on Public Policy, *J. Monet. Econ.*, Suppl. 1976, 7–64.

J. Perloff and M. Wachter, "A Production

Equity and Tradeoffs in a Tax-Based Incomes Policy

By LAURENCE S. SEIDMAN*

A tax-based incomes policy (*TIP*) is a tax incentive that tries to induce employers and/or workers to reduce wage and/or price increases. It is intended as a complement to the basic anti-inflation policy of gradual monetary deceleration and fiscal restraint. Its purpose is to reduce the capital stock and output loss that would otherwise accompany such a policy (see my 1978, 1979a articles).

This article is divided into two sections. Section I examines the capital stock and output loss under monetary deceleration that is uncomplemented by *TIP*. This loss is due primarily to "wage growth inertia," a central feature of modern capitalist economies. Such inertia provides not only the rationale for a *TIP* on wage increases; but also implies that special effort is required to assure that the design of a complete *TIP* "package" is fair to labor.

Section II analyzes the tradeoffs confronting two alternative methods of making *TIP* more equitable to labor: profit and/or price restraint insurance; and a *TIP* on price increases of the largest firms. The original *TIP*, limited to wage increases, at first glance is often perceived as unbalanced and inequitable. This section tries to contribute to the development of a balanced equitable *TIP*.

I. Wage Growth Inertia and the Output and Capital Stock Loss of Deceleration

It is not possible to have a significant permanent deceleration of price inflation without comparable deceleration of wage inflation. Costs must be covered, and unit labor cost is roughly two-thirds of value-added price. It is essential, therefore, to ask: How does a reduction in the growth rate of nominal aggregate demand, engineered by monetary deceleration and fiscal restraint, affect wage increases?

The experience of the 1975 recession provides one basis for estimating the output loss that would be required to induce the required reduction in wage inflation. The growth rate of compensation per hour declined from a peak of 11.5 percent (annual rate) in 1974:3 to a trough of 6.6 percent in 1975:4. The entire decline, however, cannot be attributed to the recession. Because the *CPI* inflation rate has some influence on wage increases—directly through cost-of-living adjustments and indirectly through its impact on expectations—the OPEC price bulge contributed to the 11.5 percent peak, and its abatement in 1975, to the 6.6 percent trough. The recession, therefore, accounted for at most a 3.5 percent decline in the wage inflation rate. (3.5 percent exceeds the estimate implied by most econometric wage equation studies.) An output loss more than twice as large as actually occurred would have been required to bring the wage inflation rate into line with the trend growth rate of labor productivity (at most, 2 percent).

How large was the output loss from the 1975 recession? Only a very rough estimate will be provided here. From 1965 to 1973, labor productivity growth averaged 1.6 percent per year in the private business sector. Assume that, had no recession occurred, the productivity growth rate would have averaged 1.2 percent per year through the decade (its actual growth rate from 1973 to 1978 was 0.8 percent). The annual growth rate of the civilian labor force averaged 2.5 percent per year from 1971 to 1979. Thus, the growth rate of "potential" (as defined here) *GNP* would have averaged 3.7 percent per year.

*Swarthmore College. This research was supported by a grant from the Employment and Training Administration of the U.S. Department of Labor.

A measure of output loss is obtained by comparing the path of actual *GNP* to the path of potential *GNP* which is projected to grow 3.7 percent per year from its level in 1971 (1971 is chosen as a benchmark because the unemployment rate averaged 5.9 percent, which is probably sustainable without raising the inflation rate). Beginning with the onset of recession in 1974, potential *GNP* exceeds actual *GNP* for the rest of the decade. The ratio of the gap to potential *GNP* is largest in 1975, 6.2 percent, and averages 3.4 percent from 1974 to 1979. In current dollars in 1980, this 3.4 percent per year average would be roughly $90 billion.

It is sometimes assumed that the output loss of the deceleration policy persists only until the unemployment rate returns to the "natural rate" (*NAIRU*). This assumption is incorrect. The unemployment rate fell to 6.0 percent in 1978, and 5.8 percent in 1979 —roughly the *NAIRU*. Yet the positive gap between potential and actual *GNP* persists through 1980. The primary reason is the investment and capital stock loss during the recession, which implies that the capital stock is lower in 1980 than it would have been had there been no recession in 1975.

If the path of "potential" gross investment (assumed to grow 3.7 percent from its 1971 benchmark) is compared to the path of actual gross investment, then a positive gap begins in 1974, and persists through 1978. The ratio of the gap to potential investment was 25.9 percent in 1975, and 13.5 percent in 1976. None of this lost capital stock has yet been recovered as of 1980 (the gap has not yet turned negative).

Thus, the use of the words "transitory" or "temporary" to characterize the output loss due to deceleration is misleading. Lost capital stock will only be recovered many years after the unemployment rate has returned to the *NAIRU*. The output loss due to the capital stock loss therefore continues over a long period.

An output and capital stock loss more than twice as large as that of the 1975 recession would be required without *TIP* to bring the current (mid-1980) wage inflation rate of 9 percent into line with the productivity growth rate. The output loss per year would therefore exceed 7 percent of potential *GNP* (roughly $180 billion in 1980), and would persist well beyond the recession due to the capital stock loss. It seems doubtful that the administrative and allocative cost of the *TIP* policy described in the next section would exceed the cost of deceleration without *TIP*. As James Tobin has observed, "It takes a heap of Harberger triangles to fill an Okun Gap."

Efficiency, however, should not be the only criterion for choosing an anti-inflation strategy. The distributional impact of each strategy must be central to the choice. Those with least skills and seniority will bear a disproportionate share of the burden of a deceleration policy uncomplemented by *TIP*. The cost of *TIP*—administrative and allocative—will be spread more evenly over the population.

It is inappropriate to argue that the more efficient policy should be chosen, and then supplemented by tax-financed transfers to achieve the desired distributional impact. Such tax-financed transfers are constrained by their own disincentive effects, are not regarded as an adequate substitute for employment by most recipients or donors, and would be limited in practice. The low-skilled worker with little seniority will bear a heavier burden under deceleration without *TIP*, given the tax transfers that could conceivably be enacted, than under deceleration with *TIP*.

II. Equity for Labor under *TIP*— The Tradeoffs

Administrative feasibility argues strongly for limiting *TIP* to the largest corporations in the economy. The largest 2,000 corporations contribute almost half the economic product, employment, and wage bill. Yet there are about 2 million corporations, 11 million sole proprietorships, and 1 million partnerships that file tax returns. Compliance and monitoring cost can therefore be greatly reduced by exempting all but the largest firms from *TIP*, and this constraint will accepted in the analysis that follows.

The "employer-penalty" *TIP* on wage increases, originally proposed by Weintraub and Wallich, should probably be the central ingredient of a *TIP* "package" (see my 1978 article). For example, when such a *W-TIP* is introduced with an initial wage guidepost of perhaps 5 percent, the base corporate tax rate for all firms (large and small) might be cut perhaps to 31 percent (to promote capital formation). For each 1 percent by which a large firm's wage increase exceeds the 5 percent standard, its tax rate might be raised 15 percent. Thus, a large firm granting 6 percent would have a tax rate of 46 percent; if 7 percent, a tax rate of 61 percent. A *W-TIP* is compatible with any base corporate tax rate, including 0 percent, should this be desired to stimulate capital formation. If the corporate income tax were converted to a value-added tax, the *TIP* surcharge could be imposed on the firm's value-added tax rate. Alternatively, the *TIP* surcharge might be imposed on the payroll tax, rather than income tax.

Would *W-TIP* be more effective than monetary deceleration in countering wage growth inertia? An important difference is in the certainty of the penalty at the time the wage decision is made. Although some economists believe that it would be "rational" for employers and workers to reduce today's wage increase if the Federal Reserve pledges to pursue monetary deceleration, many employers and employees (including economists) do not regard such behavior as "sensible." Can an employer persuade himself, and his workers, that although sales and profits are strong today, today's wage increase should be adjusted downward, because Federal Reserve policy implies that without such adjustment, sales and profits may be weak next year, and layoffs may result?

A *TIP* makes more certain, at the time of the wage decision, the penalty that will be inflicted on future after-tax profit due to a given wage settlement. Under the tax schedule just given for illustration, management and union negotiators would recognize that a 5 percent settlement implies a 31 percent tax rate for the firm, while an 8 percent settlement implies a 76 percent tax rate. Union leaders and most rank and file workers would probably realize—though they might not acknowledge it publicly—that neither their own company, nor others covered by *TIP*, would be able, or willing, to grant an 8 percent settlement, and that a strike aimed at winning 8 percent would probably be futile. If workers believed that the *TIP* "package" was grossly unfair, futile strikes might nevertheless be attempted as a form of political protest, and an expression of anger. If the measures to be discussed shortly succeed in persuading labor that *TIP* is not an unfair policy, then it seems plausible that unions would choose to avoid strikes they cannot win. Labor's strategy is generally to push management to, but not over, the brink. If *TIP* moves the brink, but is not perceived by labor as grossly inequitable, it seems possible that any rise in strike activity would be modest and temporary.

As I have discussed elsewhere (1978), it might be possible to supplement the employer-*W-TIP* with an employee-*W-TIP* that would either cut the employee tax rate at a firm granting less than the guideline, or raise the tax rate at a firm that exceeds the guideline, or both. If administrative feasibility requires limiting *TIP* to the largest firms, it might be objected that it would be unfair to restrict rewards or penalties to only half of the workers in the economy. On the other hand, it could be argued that workers at large firms are already treated differently (fringe benefits, job security), so that limiting an employee-*W-TIP* to such workers is defensible.

Whether *TIP* "works" depends on whether workers perceive it as a reasonably balanced fair policy. A wage *TIP*, by itself, would be perceived as unfair by many workers. I want to consider two alternative ways of making *TIP* more balanced: profit and/or price restraint insurance; and a *TIP* on price increases of large firms.

A. *Profit and/or Price Restraint Insurance*

The insurance strategy would attempt to respond to two legitimate concerns of labor under a wage *TIP*. First, suppose that the large corporations reduce their wage increases, but simultaneously raise their after-tax profit. Second, suppose they cut their wage increases, but price increases are not cut comparably. The aim is to design insurance that protects labor against these concerns, but at the same time satisfies the following constraints (see my 1979b article):

1) The insurance should not require measurement of the wage or price increase at individual firms (except perhaps at the largest 2,000, where wage measurement is already required by *TIP*). Exempting all but the largest firms from the *TIP* package greatly reduces compliance and monitoring cost. This constraint rules out the version of "real wage insurance" (*RWI*) proposed by the Carter Administration in 1978, because that version would have required virtually all firms to measure their own wage increase.

2) It should not try to penalize "excess" profit at an individual firm. Such an attempt would mitigate the incentive of each firm to improve efficiency, from which labor benefits.

3) It should not permanently "freeze" the distribution of income between labor and capital. The evolution of returns to factors of production at least partly in response to market forces, helps promote allocative efficiency.

4) It should not raise the tax burden on corporate income. It can be argued that a reduction of the corporate income tax would promote capital formation and allocative efficiency, from which labor ultimately benefits.

5) It should not commit the government to destabilizing automatic expenditures. An anti-inflation package should nof be obliged to pump new demand into the economy at a time when restraint is necessary.

The following two insurance policies satisfy the above constraints: Profit Restraint Insurance (*PRO*), and Price Restraint Insurance (*PRI*). My version of *PRO* has three main attributes:

1) A uniform surcharge on the largest corporations. The surcharge would be the same for all large firms covered by *TIP*. It would be imposed if the ratio of aggregate profit to aggregate labor compensation for these firms increased abnormally in a given year. A firm's own ratio would not affect its surcharge. If the base corporate tax rate were sufficiently cut when *TIP* was introduced, the *PRO* surcharge could be compatible with a reduction in the overall tax burden on corporate income.

2) No permanent constraint on income distribution. The surcharge would be imposed only in a year when the aggregate ratio increases (abnormally). If the level of the ratio remains high in the following year, the surcharge would not be continued.

3) Only an "abnormal" increase would trigger the surcharge. Normal, cyclical increases in the ratio would be estimated by econometric technique.

Given the constraints, *PRO* can only be partial insurance. It would prevent a sudden shift in after-tax income from labor to capital following *TIP*'s introduction. It would not, however, permanently prevent such a shift. Although this may seem less satisfactory to labor than complete insurance, it could be noted that labor currently possesses no profit restraint insurance.

One risk of proposing *PRO* is that, as Congress shapes the legislation, the constraints will be disregarded, and the version of *PRO* that emerges will in fact penalize a firm according to its own "excess" profit, permanently freeze the capital-labor income distribution, and permanently raise the tax burden on corporate income, with harmful consequences to the economy.

The basic idea for Price Restraint Insurance was suggested by Arthur Okun, and is similar to the "real wage insurance" proposal of the Carter Administration. My version of *PRI* has three main elements:

1) The productivity norm. A tax rebate to low- and middle-income households would be authorized when the increase in the average real wage for the economy was significantly below the increase in average labor productivity for that year. If *TIP* helps cause an initial reduction in the average wage increase to 6 percent, then a price increase of 6 percent would not authorize a rebate if productivity growth were 0 percent; but would, if productivity growth were 2 percent. Price "restraint" should therefore

be defined by the comparison of real wage growth with productivity growth.

2) The tax rebate should not depend on an individual's own wage behavior, or his firm's (as in the Carter Administration's "real wage insurance"). The aim here is to avoid measurement by firms. *PRI* is not intended as an incentive, but only as insurance.

3) Postponement with interest. To prevent a destabilizing automatic expenditure, the *PRI* tax rebate could be delayed unless the unemployment rate sufficiently exceeds the estimated *NAIRU* (perhaps 6.0 percent). Interest could accrue on the principal, and the rebate could be triggered automatically when the unemployment rate rises to the prescribed level.

As in the case of *PRO*, one risk of proposing *PRI* is that the actual legislation will disregard the constraints, so that the enacted *PRI* will ignore productivity, require wage measurement at most firms, and actually trigger the rebate when the economy requires restraint.

B. *A TIP on Price Increases of the Large Corporations*

The insurance policies just described would improve the fairness of *TIP*. They would not, however, appear to treat wages and prices symmetrically. Most would probably agree that a *TIP* on price increases would be the surest way to improve the appearance and reality of balance and equity. But is a *TIP* on price increases of the largest firms administratively feasible, and would it be reasonably equitable in its treatment of firms in different sectors?

Under the Carter Administration's wage and price standards, the Council on Wage and Price Stability obtained a measure of the average wage and price increase at each large firm willing to comply. Based on conversations with corporate managers responsible for the actual computation in several firms, my impression is that measuring the average price increase is not significantly more difficult than measuring the average compensation increase; and that neither measurement presents insurmountable obstacles for large firms, though obviously distortion is possible in both calculations.

"Perfect" measurement is not required for an adequate incentive effect. Suppose that a firm would report a 6 percent increase when auditors would have measured a 7 percent increase. Nevertheless, management would realize that if it can reduce its genuine price increase to 6 percent, it can report 5 percent and reduce its tax rate. Some slippage between reported and genuine price increases would not undermine the incentive effect of *TIP*. To keep perspective, it might be noted that such slippage is not unknown to our current tax system without *TIP*.

An incentive on price increases has been suggested by others (see Maurice Scott, Okun, Abba Lerner and D. Colander, and R. Ashley). My version of a price *TIP* would contain the following features:

1) Value-added price. One problem that has plagued previous price standards is whether a "cost pass-through" should be allowed for intermediate product purchased from other firms. *Ad hoc* adjustments have sometimes been made for sectors that experience sharp increases in raw material costs. Basing the tax incentive on the increase in value-added price is a systematic way of approaching this problem. The firm can be instructed to deflate by a price index its sales, and expense on intermediate product. Dividing sales minus intermediate product, undeflated, by sales minus intermediate product, deflated, yields value-added price.

The largest component of value-added price is unit labor cost, and it could be argued that a value-added price *TIP* could be a substitute for, not complement to, a wage *TIP*. In my view, however, more effective pressure can be exerted on collective bargaining if there is an explicit wage *TIP*, so that workers know that the company cannot escape a large tax penalty if a large wage increase is granted. Because wage growth inertia is the crux of the problem, maximum pressure should be exerted at the time of wage negotiations.

2) Sector productivity adjustment and multiyear averaging. In a given year, the standard deviation of value-added price increases exceeds the standard deviation of

wage increases. One source of price increase variation is productivity increase variation. My suggestion is that instead of a uniform price standard, Commerce Department data on sector productivity growth should be used to adjust the standard for each sector. For example, if average productivity growth for the economy is 2 percent, and the wage standard is 5 percent, then the average price standard would be 3 percent. But if sector *A* has productivity growth of 4 percent in year *t*, the price standard for any large firm in *A* would be 1 percent; if sector *B* has productivity growth of 0 percent, its price standard would be 5 percent.

The firm would not measure its own productivity increase. Standard Commerce Department data would be used to establish sector price standards, which would be given in a table in corporate tax schedule *T*. As long as each sector is sufficiently large, no single firm will have more than a trivial impact on its sector price standard, so that this procedure will have no incentive effect.

Even with this productivity adjustment, the data show that, in any one year, there can be substantial divergence between a sector's value-added price increase, and its unit labor cost increase. There is a tighter relationship, however, over a three-year period. This suggests that the firm's *P-TIP* surcharge might be based on its average price increase during the current and perhaps previous two years, and that the sector productivity adjustment might similarly be based on the same three-year period.

3) Both tax penalty and reward. The firm would be subject to an incentive over a wider range if its tax rate is cut for each point below the standard, as well as increased for each point above it. Rewarding a firm for reducing its wage increase below the wage standard might be unacceptable to labor. But a "continuous" penalty-reward *TIP* on price increases should not cause a comparable problem.

III. Conclusion

Given the large output and capital stock loss from monetary deceleration without *TIP*, and the appearance of inequity of a *TIP* limited to wage increases, careful consideration should be given to complementing a wage *TIP* with either: profit and/or price restraint insurance; or a *TIP* on price increases of large corporations. Because a price *TIP* on large firms would probably be more successful than insurance in achieving both the appearance and reality of equity, and because it would probably have less harmful side effects than any insurance policy that emerges from Congress, a price *TIP* is probably the preferred complement to a wage *TIP*, provided both are limited to large firms.

REFERENCES

R. Ashley, "Anti-Inflation Taxation Policy: An Alternative to Wage and Price Controls," *J. Macroeconomics*, Fall 1979, *1*, 417–21.

Abba P. Lerner and D. Colander, *MAP: A Market Anti-Inflation Plan*, New York 1980.

A. Okun, "The Great Stagflation Swamp," *Challenge*, Nov./Dec. 1977, *20*, 6–13.

M. F. Scott, "A Tax on Price Increases?," *Economic J.*, June 1961, *71*, 350–66.

L. Seidman, "Tax-Based Incomes Policies," *Brookings Papers*, Washington 1978, *2*, 301–48.

______, (1979a) "The Role of a Tax Based Incomes Policy," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 202–06.

______, (1979b) "*TIP*: Feasibility and Equity," *J. Post-Keynesian Econ.*, Summer 1979, *1*, 24–37.

J. Tobin, "How Dead is Keynes," *Econ. Inquiry*, Oct. 1977, *15*, 459–68.

MODELS OF LABOR MARKET EQUILIBRIUM

Implicit Contracts, Moral Hazard, and Unemployment

By SANFORD J. GROSSMAN AND OLIVER D. HART*

This paper considers a firm whose marginal (revenue) product of labor is a random variable. We derive the form of an optimal long-term contract between workers and the firm under the assumption that labor's marginal product is observed by the firm but not by the workers. We show that the existence of asymmetric information causes unemployment to be greater than in a situation where information about labor's marginal product is public, or where employment is determined in spot markets. In particular, unemployment can occur when the marginal product of labor exceeds the reservation wage.

I. The Model With One Worker

Let the firm be able to employ at most one worker, whose output is denoted by the continuous random variable $\tilde{s}$. Let $G(s)$ be the distribution function of $\tilde{s}$, and assume $G(s_0)=0$, $0<G(s)<1$ if $s_0<s<\bar{s}$, $G(\bar{s})=1$, so $s_0 \leqslant \tilde{s} \leqslant \bar{s}$. The worker can either work or not work. If he works, his utility is $U(W\text{–}R)$, and if he does not work, his utility is $U(W)$, where W is wealth. Thus his reservation wage is R. Let $V(\pi)$ denote the firm's utility of profit, and $w_u(s)$, $w_e(s)$ its payment to the worker in unemployed and employed states, respectively. We assume that the firm and the worker are risk averse, with $U''(w) \leqslant 0$, $V''(\pi)<0$.

Assume that the worker can achieve a utility level of $\bar{U}$ if he does not make a contract with this firm but goes elsewhere. We will assume that the contract is made at a time when the distribution function G is known, but before the realization of $\tilde{s}$ occurs. Consider first the case where $\tilde{s}$ is public information and where therefore the contract can be made conditional on the realization of $\tilde{s}$. Then a necessary condition for the contract to be *ex ante* Pareto optimal is that it is *ex post* Pareto optimal for each s. It follows that it is optimal to employ the worker in exactly those states where $\tilde{s} \geqslant R$. An *ex ante* optimal contract is therefore characterized by this employment rule and a choice of $w_e(\cdot)$ and $w_u(\cdot)$ to maximize

$$\text{(1a)} \qquad \int_R^{s_1} V(s-w_e(s))\,dG(s) + \int_{s_0}^{R} V(-w_u(s))\,dG(s)$$

$$\text{(1b)} \quad \text{subject to} \int_R^{s_1} U(w_e(s)-R)\,dG + \int_{s_0}^{R} U(w_u(s))\,dG \geqslant \bar{U}$$

In (1a), the expected utility of the firm is divided into its utility of profit in employment states (which is $V(s-w_e(s))$) and in unemployment states (which is $V(-w_u(s))$ since there is no output but $w_u(s)$ must be paid to the laid-off worker). Similarly, in (1b) the worker's utility in an employment state is $U(w_e(s)-R)$, and in the laid-off state is $U(w_u(s))$.

Note that the solution to the maximum problem involves setting $w_u(s)=w_u$ a constant, and the first-order condition

$$\text{(2)} \qquad \frac{V'(s-w_e(s))}{V'(-w_u)} = \frac{U'(w_e(s)-R)}{U'(w_u)}$$

*University of Pennsylvania, and University of Cambridge, respectively. Research supported by NSF grant no. 79-13429.

From (2), it is clear that since the firm is risk averse, $w_e(s)$ must depend on s, i.e., w_e is not constant.

However, if the firm was risk neutral so that $V'(\pi)$ was constant, then $w_e(s)$ would be a constant. This is the case studied by Costas Azariadis and others (see his Survey). It is clear that when the firm is risk neutral, it will not lie about s, when $\tilde{s}$ is not public information. This is because, with w_e and w_u each independent of s, the only way in which lying about s would change the wage payment is if it changes the employment status of the worker. But when V' is a constant, (2) implies that $w_e - R = w_u$. Thus the firm saves R dollars when it claims that $\tilde{s}$ is so low that there should be unemployment, but it loses s. It follows that it will never pay the firm to lay off the worker when the true s is larger than R, or to employ the worker when the true s is less than R. In other words, the firm has no incentive to lie. Hence, when the firm is risk neutral, the implicit contract model with asymmetric information gives exactly the same level of employment in each state as the implicit contract model with public information, and this is in turn the same employment level as would occur if there was a spot market in labor. A spot market in labor, of course, guarantees *ex post* Pareto optimality for each s.

It should be pointed out that many models of implicit contracts assume that the firm cannot pay the worker when he is laid off (compare Azariadis), i.e., w_u is forced to be zero. This has the consequence that the implicit contract model with public information gives different employment levels from those of a spot market model. Without modeling uncertainty about the reservation wage, setting $w_u \equiv 0$ is an artificial assumption. Furthermore, as George Akerlof and Hajime Miyazaki have pointed out, this assumption implies that there should be *less* unemployment in a contract model than in the spot market model. In particular, there will never be unemployment in states where the marginal product of labor exceeds the reservation wage.

We see then that, if either $\tilde{s}$ is public information or the firm is risk neutral, the implicit contract model cannot explain unemployment in states where the marginal product of labor is larger than the reservation wage. We now show that this possibility can be explained in the model described above if we suppose that (a) $\tilde{s}$ is observed by the firm but not by the worker; (b) the firm is risk averse.

There are two reasons why the directors of the firm and the firm's shareholders might be risk averse with respect to the firm's profit. First, if this firm's marginal product is correlated with the income shareholders get from other firms, then this firms's profit will not be a diversifiable risk. This may be of great importance in studying *macroeconomic* random shocks. Second, the fact that the firm's profitability is private information can create moral hazards which prevent shareholders and directors from diversifying this risk even if the firm's marginal product is uncorrelated with the rest of the economy.

Note that, if the firm is risk averse, the optimal contract under public information, given by (2), is no longer feasible once $\tilde{s}$ cannot be observed by the worker. For, as we have seen, $w_e(s)$ must depend on s to get optimal risk sharing between the worker and the firm. However, this will give the firm an incentive to understate s in order to reduce wage payments, for example, if $w_e(s_1) < w_e(s_2)$, where s_1, s_2 are both employment states, then the firm will claim that $s = s_1$ when $s = s_2$. Following the suggestion of Guillermo Calvo and Edmund Phelps, and Robert Hall and David Lilien, we consider now the optimal contract under the assumptions that 1) s is unobservable to the workers; 2) the *ex post* level of employment is public information. Hence, although wages cannot be made to depend directly on s, they can be made conditional on the employment level. With one worker, this means that a contract involves two wages, w_e and w_u, to be paid on employment and unemployment, respectively. Let $k \equiv w_e - w_u$. A contract must also specify an employment rule. But since nothing is observed about the state other than employment, there is no way that a contract can constrain the firm's employment choice to be any function of s which is not *ex post* optimal for the firm. That is, no matter what employment rule is specified, the firm will employ the worker iff $\tilde{s} \geqslant k$ (the firm will always find it profitable to ensure employment iff $\tilde{s} \geqslant k$, and it will do so by claiming $\tilde{s}$ is an employment state). Hence an optimal contract when $\tilde{s}$ is not public information involves choosing k and w_u to maximize

$$(3) \quad \int_k^{\bar{s}} V(s-k-w_u)dG + \int_{s_0}^{k} V(-w_u)dG$$

$$(4) \quad \text{subject to } \int_k^{\bar{s}} U(k+w_u-R)dG + \int_{s_0}^{k} U(w_u)dG \geqslant \bar{U}$$

In the next section, we prove a theorem which as a special case shows that the optimal k for (3)–(4) is larger than R. Since employment occurs only if $\tilde{s} \geqslant k$, this implies that we obtain unemployment in states for which a spot market model would not predict unemployment. Here we illustrate the basic idea. Suppose that $k=R>s_0$. This implies that the worker's net income is constant in employment and unemployment states. (If the worker is paid his reservation wage for working, then he is indifferent between working and not working.) Thus the worker is locally risk neutral; that is, even though he may be very risk averse, he will always be better off with a sufficiently small risky gamble if the expected payoff of the gamble is positive. If we add the expected net income of the worker EI, to the expected net profit of the firm $E\pi$, we get

$$(5) \qquad E\pi + EI = \int_k^{\bar{s}} (s-R)dG(s)$$

Clearly (5) is maximized at $k=R$. Thus a small increase in k at $k=R$ will have only a second-order effect on $E\pi+EI$. The fact that, at $k=R$, the firm is bearing all the risk suggests that an appropriate increase in risk sharing will improve welfare. Suppose that we increase k a little from $k=R$ and decrease w_u a little in such a way as to keep $E\pi$ constant (this means increasing $(k+w_u)$). Then since $s-(k+w_u)>-w_u$ in employment states (i.e., the firm is better off in employment states than in unemployment states), this results in a transfer of income from good (employment) states to bad (unemployment) states. Hence the firm faces a less risky profit stream. It follows that if the firm is strictly risk averse it will accept a decrease in $E\pi$ in exchange for this decrease in risk. Therefore choose the small increase in k and reduction in w_u so that the firm is made better off and $E\pi$ falls. Then, from (5), EI goes up and hence, since the worker is locally risk neutral at $k=R$, he is also made better off. This shows that $k>R$ is Pareto superior to $k=R$.

The above argument assumes that $R>s_0$. A more extreme case occurs when $R \leqslant s_0$, i.e. $R \leqslant \tilde{s}$ for all $\tilde{s}$. Then, under public information or spot markets, there is full employment; that is, there are no layoff states and risk sharing is achieved entirely by variations in $w_e(s)$. If $\tilde{s}$ is not publicly observable, however, this contract is not feasible since, for moral hazard reasons, variations in w are now possible only if they are accompanied by movements in employment. Hence full employment implies a constant wage, which means that the firm bears all the risk and the worker bears none. In general (although not always), the firm and the worker will prefer a contract which sacrifices employment in low productivity states where $\tilde{s}>R$, but which permits some risk sharing through a variable wage. The theorem of the next section covers this case as well as the case $R>s_0$.

II. A Proof that Moral Hazard Leads to Underemployment

In this section, we prove the main theorem for the case of many workers. Suppose there are n workers, each with the same utility function as the single worker in the last section. Let the firm's production function be $q=f(s, l)$, where l is the number of employed workers. To model the idea that shifts in s represent marginal product shifts, we assume that, for any l,

$$(6a) \qquad f(s,l)-f(s,l-1)$$

is an increasing function of s, and

$$\text{(6b)} \qquad f(s,l) \text{ is increasing in } s \text{ and } l, \quad f(s,0)=0, f \text{ is concave in } l$$

If $\tilde{s}$ was public information, then an employment policy would involve dividing up the interval $[s_0, \bar{s}]$ into subintervals $[s_0, \bar{s}] = [s_0, s_1] \cup [s_1, s_2] \cup [s_2, s_3] \cup \ldots \cup [s_n, \bar{s}]$ such that, if $\tilde{s} \in [s_i, s_{i+1}]$, then exactly i workers are employed; for example, if $\tilde{s} \in [s_n, \bar{s}]$, then all n workers are employed. These regions can be found by solving the equations $f(\hat{s}_1, 1) = R$, $f(\hat{s}_2, 2) - f(\hat{s}_2, 1) = R, \ldots, f(\hat{s}_n, n) - f(\hat{s}_n, n-1) = R$, and then setting $s_i = \hat{s}_i$ if $s_0 < \hat{s}_i < \bar{s}$, $s_i = s_0$ if $\hat{s}_i < s_0$, $s_i = \bar{s}$ if $\hat{s}_i > \bar{s}$. This ensures that the correct number of workers will be hired given that each worker must be paid his reservation wage R. To summarize, under public information (or under spot market wage contracts), i workers will be employed in state s iff $s \in [s_i, s_{i+1}]$, where

$$\text{(7)} \qquad f(s_i, i) - f(s_i, i-1) = R$$

If $\tilde{s}$ is private information and contracts are made, then these can be conditioned only on the level of employment. A contract involves a set of wages $w_0, w_1, \ldots, w_n$, where w_i is the *total wage bill* of the firm if it hires i workers. The firm's total wage bill if it hires i workers is

$$\text{(8)} \qquad w_i = i x_i + (n-i) y_i$$

where x_i is the payment to an employed worker and y_i is the payment to a laid-off worker. As explained in the last section, since employment is the only public variable, given $\{x_i, y_i\}$, the only incentive-compatible employment contract is one where, in each state s, the firm will *ex post* decide on an employment level to maximize profit subject to the conditions of the wage contract. In state s, the firm will choose l to maximize $f(s, l) - w_l$. Thus, given the wage contract $\{w_i\}_i$, it is possible to find the regions in which the firm will employ exactly $0, 1, 2, \ldots, n$ workers.

Let $l(s)$ maximize $f(s, l) - w_l$. (6a) implies that $l(s)$ is increasing in s. Therefore an employment policy can again be represented by a division of the interval $[s_0, \bar{s}]$ into subintervals $[s_0, \bar{s}] = [s_0, k_1] \cup [k_1, k_2] \cup [k_2, k_3] \cup \ldots \cup [k_n, \bar{s}]$ such that if $\tilde{s} \in [k_i, k_{i+1}]$ then exactly i workers are employed, where $s_0 \equiv k_0$, $\bar{s} \equiv k_{n+1}$. It turns out to be more convenient to regard $k_0, \ldots, k_{n+1}$ rather than $w_0, \ldots, w_n$ as control variables. Given an employment policy $[s_0, \bar{s}] = [s_0, k_1] \cup [k_1, k_2] \cup \ldots \cup [k_n, \bar{s}]$, the corresponding wage bills $w_0, w_1, \ldots, w_n$ must satisfy

$$\text{(9)} \qquad f(k_i, i) - f(k_i, i-1) = w_i - w_{i-1} \qquad i = 1, \ldots, n$$

A Pareto optimal contract involves wages $\{w_i\}$, $\{x_i, y_i\}$ and employment rules $\{k_i\}$, where $s_0 \equiv k_0 \leqslant k_1 \leqslant \ldots \leqslant k_n \leqslant k_{n+1} \equiv \bar{s}$, such that

$$\text{(10a)} \qquad \sum_{i=0}^{n} \int_{k_i}^{k_{i+1}} V(f(s,i) - w_i)\, dG(s)$$

is maximized subject to (8) and (9) and

$$\text{(10b)} \qquad \sum_{i=0}^{n} \int_{k_i}^{k_{i+1}} \left[U(x_i - R)\frac{i}{n} + U(y_i)\left(\frac{n-i}{n}\right) \right] dG(s) \geqslant \bar{U}$$

In (10b), we have assumed that when a layoff occurs a given worker is drawn at random from the employment pool n. The problem in (10) can be simplified by noting that the risk of being the one chosen to be unemployed out of the labor pool is diversifiable by the employer. The employer is risk averse about the size of the whole wage bill; he is indifferent as to how it is divided between the workers (see Akerlof and Miyazaki). Thus it is optimal to set $x_i - R = y_i$ so that workers are insured about whether they are the unlucky ones chosen to be laid off, given that $n-i$ are laid off. Equation (8) implies that the net income to a given worker is $y_i = x_i - R = (w_i - iR)/n$. Thus, (10b) becomes

$$\text{(10c)} \qquad \sum_{i=0}^{n} \int_{k_i}^{k_{i+1}} U\left(\frac{w_i - iR}{n}\right) dG \geqslant \bar{U}$$

Our object is to show that there is more unemployment when $\bar{s}$ is private information than would occur if it was public (or equivalently if trades were forced to take place on spot markets). What this means is that, in a given state s, if spot markets yielded i employed workers, then implicit contracts will yield fewer than i employed. Equivalently, given that, for $s \in [s_i, s_{i+1}]$, exactly i workers are employed in a spot market where s_i is given by (7) then $k_i \geqslant s_i$ for all $i = 1, \dots, n$ since this means that a higher s must be realized before the firm decides to employ i workers. (Recall that i workers are employed iff $s \in [k_i, k_{i+1}]$.) This is our theorem:

THEOREM : *Assume* $V''(\pi) < 0$ *and let* $\{k_i, w_i\}$ *maximize* (10a) *subject to* (9), (10c) *and* $s_0 \equiv k_0 \leqslant k_1 \leqslant \dots \leqslant k_n \leqslant k_{n+1} \equiv \bar{s}$. *Then* $k_i \geqslant s_i$ *for* $i = 1, \dots, n$, *where* s_i *is the spot market cut-off employment state given by* (7), *and* $k_i > s_i$ *as long as* $\bar{s} > s_i > s_0$, i.e., *as long as i lies between the minimum and maximum employment levels under public information.*

PROOF:

We prove the result first for the case where $s_0 \equiv k_0 < k_1 < \dots < k_{n+1} \equiv \bar{s}$ at the optimum. In this case, we can invert (9) to solve for k_i (denote the solution by $k_i = h_i(w_i - w_{i-1})$) and regard $w_0, \dots, w_n$ as the control variables. Let $\lambda \geqslant 0$ be the Lagrange multiplier for the constraint (10c). Then maximizing (10a) subject to (10c) yields the first-order conditions

$$\text{(11a)} \quad \bar{\alpha}_0 - \alpha_0 b_0 = \delta_1 \Delta U_1$$

$$\text{(11b)} \quad \bar{\alpha}_i - \alpha_i b_i = \delta_{i+1} \Delta U_{i+1} - \delta_i \Delta U_i \qquad i = 1, 2, \dots, n-1$$

$$\text{(11c)} \quad \bar{\alpha}_n - \alpha_n b_n = -\delta_n \Delta U_n$$

where

$$\text{(12)} \quad \bar{\alpha}_i \equiv \int_{k_i}^{k_{i+1}} V'(f(s,i) - w_i)\, dG(s)$$

$$\text{(13)} \quad \alpha_i \equiv G(k_{i+1}) - G(k_i), \; b_i = \frac{\lambda}{n} U'\left(\frac{w_i - iR}{n}\right)$$

$$\text{(14)} \quad \delta_i \equiv \lambda h_i'(w_i - w_{i-1}) G'(k_i)$$

$$\text{(15)} \quad \Delta U_i \equiv U\left(\frac{w_i - iR}{n}\right) - U\left(\frac{w_{i-1} - (i-1)R}{n}\right)$$

Clearly, (11) implies that $\lambda > 0$. Note that by the intermediate value theorem, there exists an $\bar{s}_i \in (k_i, k_{i+1})$ such that, if $g_i \equiv V'(f(\bar{s}_i, i) - w_i)$, then

$$\text{(16)} \quad \alpha_i g_i = \bar{\alpha}_i$$

Note also that

$$\text{(17)} \quad g_0 > g_1 > \dots > g_n$$

This is because $V'' < 0$ and $f(\bar{s}_i, i) - w_i > f(\bar{s}_i, i-1) - w_{i-1} > f(\bar{s}_{i-1}, i-1) - w_{i-1}$. The first inequality obtains because by construction the firm is better off employing i workers when $s \in (k_i, k_{i+1})$ than $i-1$ workers. The second inequality follows from the assumption that $f(s, l)$ is increasing in s. Another important fact is that, by the strict concavity of $U(\cdot)$, $b_i < b_{i-1}$ iff $w_i > w_{i-1} + R$. Hence

$$\text{(18)} \quad \Delta U_i > 0 \text{ iff } b_i < b_{i-1}$$

Our object is to show that if s_i solves (7), then $k_i > s_i$ $i = 1, \dots, n$. Recalling that $h_i(\cdot)$ is the solution to (9) for k_i, note that $k_i > s_i$ is equivalent to $h_i(w_i - w_{i-1}) > h_i(R)$. Since h_i is increasing by (6a), this is equivalent to $w_i - w_{i-1} > R$. From (15), this is equivalent to $\Delta U_i > 0$, which from (18) is equivalent to $b_i < b_{i-1}$ for all i. We now prove the last inequality by contradiction.

Suppose $b_0 > b_1 > \dots > b_{i-1} \leqslant b_i$, so that i is the first index when $b_i < b_{i-1}$ is contradicted. From (18),

$$\text{(19)} \quad \Delta U_i \leqslant 0 \text{ and } \Delta U_{i-1} > 0$$

Hence, using (16), (11b), and the facts that $\alpha_i > 0$ and $\delta_i > 0$, yields

$$\text{(20)} \quad g_{i-1} \leqslant b_{i-1}$$

This is true even if $i = 1$ as can be verified from (11a). Hence, by (17),

$$g_i < g_{i-1} \leqslant b_{i-1} \leqslant b_i \tag{21}$$

So $g_i < b_i$. Hence, by (11b), $\delta_{i+1}\Delta U_{i+1} - \delta_i \Delta U_i < 0$, and so, by (19), $\Delta U_{i+1} < 0$. Therefore, by (18), $b_i < b_{i+1}$. This shows that if $b_{i-1} \leqslant b_i$, then $g_{i+1} < g_i < b_i < b_{i+1}$, where we are using (17). Repeating this argument with $i+1$ replacing i yields eventually $g_n < b_{n-1} < b_n$. Hence, by (18), $\Delta U_n < 0$ and $g_n < b_n$. This contradicts (11c) and (16).

We have established the theorem for the case $k_0 < k_1 < \ldots < k_{n+1}$. We now sketch the extension when some of the k_i are equal. Let $L_0, \ldots, L_J$ be the distinct employment levels corresponding to the k_i, i.e., $k_{L_0} < \ldots < k_{L_J}$. Then the above argument shows that $(w_i - L_i R)$ is increasing in i. The next step is to show that $L_{i+1} = L_i + 1$. This is proved by showing that if $L_{i+1} - 1 = L > L_i$, then 1) setting $w_L = w_{L_{i+1}} - R$ makes the workers and firm better off unless $k_{L_{i+1}} = s_{L_{i+1}}$; 2) if $k_{L_{i+1}} = s_{L_{i+1}}$, then the left-hand side of (11b) > the right-hand side when $i = L$, which means that a small reduction in w_L below $(w_{L_{i+1}} - R)$ increases welfare. The final step is to show that the minimum (maximum) employment level under asymmetric information is no greater than (is the same as) that under public information. If not, a Pareto improvement could be achieved by setting $w_i = w_{L_J} + (i - L_J)R$ for $i > L_J$, $w_i = w_{L_0} + (i - L_0)R$ for $i < L_0$.

Remark 1: The proof of the theorem establishes two further points. First, each worker's utility is an increasing function of s, the state of the world. Since the firm's profit is also increasing in s, this means that the optimal contract provides co-insurance between the firm and the workers. Secondly, given the size of the labor pool n, the maximum employment level under asymmetric information, that is, the employment level in the best state $s = \bar{s}$, is the same as that under public information. Since the theorem tells us that employment in the worst state $s = s_0$ is generally lower under asymmetric information than under public information, we may conclude that the *variability* of employment is greater under asymmetric information than under public information.

Remark 2: Note that we have taken the size of the overall labor pool n as exogenous. If this is a general equilibrium model and n is the total number of workers per firm, then *ex ante* market clearing will imply that $\bar{U}$ adjusts so that the firm finds it optimal to make contracts with n workers. More generally, it is possible to extend the above theorem to show that unemployment will be greater under asymmetric information than under public information even taking into account the fact that the firm may choose different values of n under public and asymmetric information.

III. Conclusions

One reason for studying implicit contract models is the vague empirical observation that workers are sometimes laid off in states in which their marginal product exceeds their reservation wage. This suggests that the conditions for *ex post* productive efficiency are not always met, which in turn raises the possibility that equilibrium is not brought about by the clearing of spot demands and supplies.

The earliest implicit contract models assumed no moral hazard, so there was no tradeoff between insurance and *ex post* efficiency. In models where the risk faced by the firm is purely idiosyncratic and diversifiable, the firm completely insures the worker by giving him a constant wage. This means the firm has no incentive to lie about the state of the world. Once we consider shocks which are not completely diversifiable, the desire for risk sharing by firms and workers makes it optimal to have a variable wage. Following Calvo and Phelps, and Hall and Lilien, we have assumed that the level of employment is the only public variable on which the wage can be conditioned. Then the only way to get the right degree of risk sharing (i.e., wage variability) is to have "excessive" employment variations. As a result, there will be more unemployment when information is asymmetric than would occur with spot markets or in the absence of moral hazard. This result is rather different from that of Hall and Lilien. They emphasize that it is *feasible* to induce the same employment level when marginal products are not public as when they are, by making the wage bill a function of the employment level. To do so, the firms's wage bill would have to be a linear function of the level of employment with the cost of employing another worker equal to the worker's reservation wage. However, we have shown that when the firm is risk averse, this leads to suboptimal risk sharing.

As well as showing that unemployment may occur when the marginal product of labor exceeds the reservation wage, our model can also explain another much believed "observation" about the labor market —that the market is often in a state where workers would like to supply more labor at going wages than firms permit them to; that is, workers are "rationed." The proof of our main theorem shows that an optimal contract will have the property that $w_i - w_{i-1} > R$, that is, the difference between the wage bill when i workers are employed and when $(i-1)$ are employed exceeds the reservation wage. But this is precisely the statement that an unemployed worker would choose to work rather than remain unemployed.

All variables in this paper are in real (not nominal) terms. If the money supply or the price level is public information, then our model cannot explain why nominal shocks should have real effects. Further, we have assumed that workers and firms do not observe the real shocks in the economy which may provide information about the marginal product in their firm. To the extent that nondiversifiable risks resulting from macro-economic shocks are significant, it is important to model the ability to condition wage and employment contracts on economy-wide variables, like the unemployment rate. We hope to do this in future work.

REFERENCES

G. Akerlof and H. Miyazaki, "The Implicit Contract Theory of Unemployment Meets the Wage Bill Argument," *Rev. Econ. Stud.*, Jan. 1980, *48*, 321–38.

C. Azariadis, "Implicit Contracts and Related Theorems: A Survey," working paper no. 79-17, department of economics, Univ. Pennsylvania.

G. A. Calvo and E. S. Phelps, "Employment Contingent Wage Contracts," *J. Monet. Econ.* Suppl. 1977, 160–168.

R. E. Hall and D. M. Lilien, "Efficient Wage Bargains Under Uncertain Supply and Demand," *Amer. Econ. Rev.* Dec. 1979, *69*, 868–79.

Contractual Models of the Labor Market

By BENGT HOLMSTROM*

In this paper I discuss some approaches to modelling labor markets as contractually mediated. The central thesis of such models is that due to market imperfections—absence of contingent claims for labor and income—wage mediated auction markets will in general not be sustainable. There will be an opportunity for firms and workers to make a joint long-term contract which improves the welfare of both, and causes the auction market to collapse. This view, first suggested by implicit contract theory (see Costas Azariadis; Martin Baily), opens up quite new perspectives on the operation of labor markets. Most importantly, one finds that in the contractual paradigm wage and marginal product may differ so that what appears to be disequilibrium in the short run may be a consistent equilibrium in the long run. Whether or not this can account for involuntary unemployment I will comment upon later. It relates to my main concern, which is with the extent to which complex contingent contracts can be enforced, and the implications this has on the nature of equilibrium.

I will start by surveying the main results of implicit contract theory in the light of a rather general model of contractual equilibrium (Section I). Weaknesses of implicit contract theory will lead us to question the assumption of enforceable state-contingent contracts. Section II looks at the alternative of non-contingent fix-wage contracts whereas Section III presents a simple analysis of reputation as a means of enforcing more complex contingent contracts.

I. Implicit Contracts

Since it has been widely believed that implicit contracts can emerge only in markets where it is costly to move, let me start by showing that this is an artifact of the commonly used one-period model. Assume initially that the labor market is cleared through sequential wage auctions and consider for simplicity two periods only. Current wage is w_0 and next period wage w_1 is uncertain with known distribution $G(w_1)$. A worker's expected utility of participating in the market is assumed to be $u(w_0)+\int u(w_1)dG(w_1)$, where $u(.)$ is an atemporal risk averse utility function. Firms are assumed risk neutral. The expected wage bill for one worker is $w_0+\int w_1 dG(w_1)$.

The claim is that firms can depart from the auction outcome to the benefit of both parties. To show that, consider a contract which pays the worker w' in the first period and guarantees w' in the second period as well. If market wage in the second period exceeds w', the firm has to follow suit or else the worker quits so the contract is called off. Choose w' so that

$$(1)\qquad u(w_0)+\int_0^{w'} u(w_1)dG(w_1) = u(w')\left(1+\int_0^{w'} dG(w_1)\right)$$

that is, so that the contract offers the worker the same prospect as the auction market. Dividing (1) by $(1+\int_0^{w'}dG(w_1))$ and applying Jensen's inequality, gives

$$(2)\qquad w_0+\int_0^{w'} w_1 dG(w_1) > w'\left(1+\int_0^{w'} dG(w_1)\right)$$

establishing the claim.

Note that the role of two periods is to allow the firm to collect a premium in the first period $(w_0-w'>0)$ for the insurance it provides in the second period. The suggested contract looks like an option and (2) indicates that the selling price includes a risk premium which makes the contract

*Associate professor, J. L. Kellogg School of Management, Northwestern University. I would like to thank Dale Mortensen and Edward Prescott for comments on an earlier manuscript.

favorable to the firm. However, the argument can be amended for risk averse firms. Also, identifying the worker's utility function is inessential. What is essential though is that beliefs are not too dispersed and, most importantly, that the firm will not lay off or exchange the worker when $w_1 < w'$. This last point I will return to.

Once all firms recognize the value of long-term contracts, wage auctions get replaced by contractual auctions in which the market will be equilibrated through the expected utility contracts offer. Since the market still is incomplete there will be a need to reopen it each period. What is envisioned to happen in future markets will influence current contract design and vice versa so the natural notion of equilibrium is one based on rational expectations. One can show such an equilibrium to exist. It is closely related to the pioneering models of Roy Radner and Oliver Hart, with the distinction that here securities are created endogenously by firms, and by assumption can be bought only through labor attachment at the respective firm.

Let me now turn to some properties of an equilibrium in implicit contracts, basing my discussion on an explicit treatment of a two-period version of the model sketched above (see my earlier paper). With risk-neutral firms and homogeneous labor, there will be a unique optimal state-contingent contract that firms offer to their workers. If firms differ in their risks, contracts will differ and different firms may therefore pay different wages. In general contracts will involve layoffs in bad states and if the market turns favorable some workers may quit. Wagewise, an optimal contract will look like the earlier described option. As long as market forces do not push up equilibrium expected utility levels, retained workers will enjoy a constant wage, but with increased labor demand, wages will rise (due to legal constraints on involuntary servitude). It should be stressed that the implied downward rigidity of wages relate to individual contracts rather than aggregate wage levels. There is no presumption that new generations will receive the same wage as old ones—in general they will not—and a laid-off worker who has to find a job elsewhere is normally forced to take a wage cut. Thus, optimal contracting creates endogenously seniority classes within an otherwise homogenous labor force. Therefore, in aggregate, the wage level will be flexible both up and down, but at a more sluggish pace than wage auctions would imply. This sluggishness is further increased by rights to be recalled (at previous wage) before any new workers are employed; a provision which will be part of an optimal contract.

Regarding the employment part of contracts I note that the model accomodates both quits and layoffs. Layoffs will occur for the same reason as in Azariadis' original treatment, namely the outside opportunity a worker has. But somewhat more acceptably this outside opportunity could be another firm rather than an exogenous benefit (for example, household income). Essential for understanding the determination of employment is the fact that since the implicit contract is *ex ante* efficient and state-contingent it is also *ex post* efficient (as a necessary condition for *ex ante* efficiency). It follows immediately that a worker who can produce more within the firm than outside will not be laid off. A more specific condition is obtained by equating the marginal rates of substitution between wage and employment probability (defined as the percentage of workers retained), which gives

$$pf'(nr) = w' - \frac{u(w') - u(w_1)}{u'(w')} \tag{3}$$

where p = output price (random), $f(.)$ = production function, n = labor pool, r = proportion retained, and the other variables have been defined earlier. The right-hand side is decreasing in w' and achieves its maximum for $w' = w_1$ (since $w' > w_1$ is required or else the worker quits). Thus, we find that actually labor will be retained not only beyond the point where marginal product equals contract wage but market wage! Two important conclusions follow: wage and marginal product generally differ and a divergence of the two does not signal disequilibrium (which casts some doubt on recent fix-price modelling); and, secondly, there will be less rather than more unemployment in the contractual model compared to wage auctions.

The last conclusion has been viewed as a failure of implicit contract theory to explain involuntary unemployment, and if one defines involuntary unemployment as unexploited opportunities to trade, indeed it is. However, it is not clear to what extent opportunities are left unexploited in the real economy when one looks at the labor market in isolation; what we observe taking place could be consistent with (3), with inefficiencies being due to faults in the coordination between product and labor markets instead.

In order to alter the conclusion about the level of unemployment it is necessary to bring in some elements of incomplete or asymmetric information (otherwise contracts will be *ex post* optimal). More importantly, such modifications will help to patch a logical inconsistency of implicit contract theory, namely the lack of severance payments with fully insured income. One simple change in informational assumptions is that w_1 cannot be observed by the firm and it will instead have to act on the conditional expectation $E(w_1|p)$. Then it is clear that, even if it wanted to, the firm could not guarantee the worker a constant income, and in some states (recessions) the level of unemployment would fall below that of an auction market (the rule would be as (3) with $Eu(w_1|p)$ replacing $u(w_1)$). Another change, explaining incomplete severance payments, would take note of the fact that laid off workers would have no incentive to search if their income was fully insured. With search, layoffs would result in some unemployment rather than pure transfers as in the simple model.

The third change, which I will pay more attention to, calls into question the firms assumed honesty. Generally, it is hard (for a single worker at least) to observe the marginal product of the firm and hence see if the required rule (3) is being followed. It appears therefore tempting for the firm to deviate from this behavior and insofar that this can be expected to happen, implicit contracts are rendered infeasible. Though I will indicate in the last section how a concern for reputation in the labor market may induce the firm to behave as if an implicit contract was written and honored, the doubt about enforceability of contracts prompts us to look at alternative contracts, which are not contingent on states that cannot be observed.

II. Fix-Price Contracts

The simplest non-state-contingent contract is one in which the firm guarantees the worker a nominal wage, but adjusts employment at its own discretion. Let me first analyze whether such contracts can be expected to arise endogenously as a mutually beneficial arrangement compared to the wage auction outcome. If so, we would again have an explanation for the breakdown of wage auctions and have a legitimate reason to study fix-wage contracts further. I look at an example only.

Consider an industry in which there is a large number of identical risk neutral firms characterized by the production function $f(n)$, $f'(1)=1$. The output price fluctuates randomly between 1 and $p\varepsilon(0,1)$, which both occur with equal frequency. Workers supply their unit of labor inelastically and have an atemporal utility function $u(w)$. The number of workers per firm (appropriately scaled) is $n=1$. In a wage auction therefore, wage will equal output price (1 or p). Expected utility from such a market is $Eu=1/2(u(1)+u(p))$, and expected profit $E\pi=(1+p)(f(1)-1)$.

Suppose now a firm would offer its workers a fixed wage w and lay off the unprofitable ones when output price is p. Letting r be the number (and proportion) retained, we have

$$pf'(r)=w \tag{4}$$

In order for this contract to be favorable (compared to the auction outcome) both to the worker and the firm we must have

$$E\bar{u}\equiv\tfrac{1}{2}(1+r)u(w)+\tfrac{1}{2}(1-r)u(p)\geqslant Eu \tag{5}$$

$$E\bar{\pi}\equiv\tfrac{1}{2}(f(1)-w)+\tfrac{1}{2}(pf(r)-wr)\geqslant E\pi \tag{6}$$

In (5) I use the assumption that the firm is small so a laid-off worker can take a job in the spot market at wage p. Rewriting (5) and (6),

$$(7)\qquad r(u(w)-u(p)) \geqslant u(1)-u(w)$$

$$(8)\qquad (1+p)/p-f'(r)(1+r) \geqslant f(1)-f(r)$$

It is easy to see that (4), (7), and (8) can be simultaneously satisfied; for instance, if $w \leqslant (1+p)/2$, then (4) implies that (6) (hence (8)) is satisfied for all f, and we can take f and u so that (7) holds. To understand what factors determine when (4)–(6) will hold, we can look at comparative statics. Assume that the equations above hold. Then they will remain intact either when the absolute risk aversion of u is increased, or if f becomes more kinked in the sense that $f'(r)$ and $f(1)-f(r)$ decrease (f becomes flatter between r and 1 and steeper between 0 and r). These conditions accord with intuition, since it is easy to understand (and check from (4)–(6)) that when the worker is risk neutral or when the production function exhibits constant returns to scale, there can be no gains to a fix-wage contract. Regarding changes in p, the effect is ambiguous. Note, however, that if there is a firm for which productivity does not fluctuate at all, it will always pay this firm to offer a fix-wage contract.

The assumption that the market is frictionless is rather unfavorable for a fix-wage scheme. Robert Gordon was first to suggest that a fix-wage scheme may be a response to the temptation a firm would have to lower the wage (claiming low marginal product) at will. Such behavior is, of course, limited by the worker's outside opportunity, but with costly labor mobility the firm's opportunity to exploit workers would quickly become rather substantial. My intention with the model above was to show that even without such transaction costs one can make a case for a fix-wage arrangement, and to indicate what factors will influence the benefits thereof.

In general, fix-wage arrangements will imply unemployment in excess of voluntary levels. Yet it may be an outcome of equilibrium and moreover an efficient arrangement subject to informational constraints, since as recent work on efficiency under asymmetric information emphasizes, it is false to apply standard *ex post* efficiency measures in the type of situation described.

The incomplete information paradigm has some apparent advantages over implicit contract theory. Wages are rigid without the unrealistic assumption of firm risk neutrality. With risk neutrality one is left wondering why a firm does not pay severance to laid off workers and thereby insure income (the model above begs the same question; a justification would require a state in which severance is infeasible). Indeed, the role of unemployment benefits is quite unclear—why do firms not provide those privately? When wages are fixed due to incentive considerations, firms can be assumed risk averse, explaining incomplete severance and the incentive to pool risks through jointly paid unemployment benefits. One may also expect that optimal levels of such benefits, the degree to which they should be experience rated, and any additional severance payments could be determined. These questions appear fruitful for future research.

However, one should keep in mind that, if there would be large gains to state contingent contracts, we would expect to see proxies for those states enter into contracts (a host of public data would be available for that) or observe direct monitoring of requisite states. But we do not, and before we understand why, enthusiasm over the incomplete information paradigm, in particular its implication for involuntary unemployment, should be controlled.

III. Reputation

I turn to the final point: can a concern for reputation lead the firm to act as if it honored an implicit contract? As we know from the growing literature on incentives in agencies (and more generally from the theory of repeated games), multiperiod considerations will allow a rather richer set of opportunities to combat averse effects of informational asymmetries. To illustrate how, I will again use a simple example.

Consider a firm operating in an economy that can be in one of two states $s=0$ or $s=1$. The firm's revenue function is $(1-\alpha s)\times f(n)$, where n is labor input. A suggested interpretation is that $s=1$ corresponds to a recession and the parameter α indicates how sensitive the firm is to the recession. The sensitivity parameter stays fixed over time. It is initially unknown to workers, and this creates the opportunity for the firm to build a reputation by signalling the value of α through layoff behavior.

Workers work only in one period for which they sign a fix-wage contract as in the previous section. The wage demand will depend on their assessment of the probability of layoff, labelled as the firms reputation $\rho \varepsilon (0,1)$. In order to attract workers the firm has to pay $w(\rho)$, which is defined from the expected utility expression:

$$(9)\qquad u(w(\rho))(p+(1-p)\rho) + u(\bar{w})(1-p)(1-\rho)=v$$

where $\bar{w}$ is income as unemployed, $p=Pr(s=0)$, and v is the expected utility offered by the market (assumed constant over time). Reputation ρ is a function of how the firm behaved in the previous recession. If a percentage r was laid off, this will be interpreted as indicating that the firm's type is $\alpha(r)$ (a signalling function to be determined in equilibrium), which in turn will give a prediction for how the firm will behave in the next recession $\rho(\alpha)$. Since α does not change over time in the simple version presented here, equilibrium will have $\rho(\alpha(r))=r$, that is, the firm will lay off the same amount in each recession and the predictions obtained by assuming that the firm will repeat its behavior will become self-fulfilling in equilibrium. Therefore, I take reputation formation to progress as: $\rho_t=\rho_{t-1}$, if $s_t=0$, $\rho_t=r_t$ if $s_t=1$, where t is a time index.

Let the firm's discount factor be δ, and $V(\rho)$ be the optimal discounted expected profit function given current reputation ρ. Then

$$(10)\quad V(\rho)=\max_{n,r}\{(f(n)-w(\rho)n)p + ((1-\alpha)f(nr)-w(\rho)nr) \times(1-p)+\delta V(\rho)p + \delta V(r)(1-p)\}$$

Here the firm is viewed as deciding the proportion retained r in case of a recession at the time it hires labor. Note that $\pi(n,w,r)\equiv(f(n)-wn)p+((1-\alpha)f(nr)-wnr)(1-p)$, is the one-period expected profit and that an implicit contract that is optimal solves $max_{n,r}\pi(n,w(r),r)$.

In a stationary state we want $r=\rho^*$ to solve (10), were ρ^* is the stationary retention value. Thus,

$$(11)\quad V(\rho^*)=\frac{1}{1-\delta}\pi(n^*,w(\rho^*),\rho^*)$$

where $n^*=arg_n max \pi(n,w(\rho^*),\rho^*)$. For ρ^* we get a condition by noting that it must not benefit the firm to move to any other stationary layoff policy. Thus,

$$(12)\quad \rho^*=\underset{r}{argmax}\left\{(1-\alpha)f(n^*r) - w(\rho^*)n^*r+\frac{\delta}{1-\delta}\pi(n^*,w(r),r)\right\}$$

This is the main equation of interest. Myopic behavior (as assumed in the previous section) results if $\delta=0$. Then layoffs will equate current marginal product to wage. But when $\delta>0$, so that there is a concern for the future, reputation will force the firm to set layoffs below the myopically optimal value; $((d/dr)\pi(n^*,w(r),r)>0$ for myopic r). Thus, reputation will provide for increased employment insurance. On the other hand, unless $\delta=1$, ρ^* will not be as high as in an implicit contract (which just maximizes $\pi(n^*,w(r),r)$ over r), since there is a cost of signalling through the excess employment. The optimal level of retained workers ρ^* will therefore lie between the myopic solution and the implicit contract solution, and be closer to the latter the closer δ is to 1 (the case of no discounting).

The paradigm suggested above is, of course, extremely simple and stylized, but I take it to indicate that reputation may have the power to enforce implicit contracts (or nearly so) even when the worker cannot verify the state of nature. This deserves further study. One would hope to learn how sophisticated the contingencies can be in reputation contracts, what influence the frequency of events has on reputation formation and could wage variation (when desirable) be supported by reputation as well. Intuition suggests that only rather simple, regularly occuring contingencies can be included in reputation contracts and that wages cannot be varied in such contracts because the wage part of a contract is a zero-sum game whereas the employment part is not, but the art of modelling reputation is still too primitive to confirm this intuition.

IV. Concluding Remarks

Implicit contract theory has been influential in suggesting a contractual view of labor markets, which carries the promise of a much improved understanding of how labor markets operate. The theory itself is incomplete at some crucial points. The main question taken up here concerns the enforceability of state-contingent contracts. On one hand, if we abandon the assumption that finely state-contingent contracts can be enforced, it leads to a fix-wage model which at least casually looking displays more reasonable features than the implicit contracts. On the other hand, the previous section indicates that a concern for reputation leads to behavior which appears as if implicit contracts were enforced. A consolidate view is that wages will be downward rigid largely due to enforcement problems (rather than risk sharing), whereas employment rules will, at least to some degree, reflect a concern for reputation and therefore come closer to what would be implied by implicit contracts, perhaps close enough to explain why more complex contingent contracts that would be feasible to write are not written.

REFERENCES

C. Azariadis, "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–1202.

M. N. Baily, "Wages and Employment under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

R. J. Gordon, "Recent Developments in the Theory of Inflation and Unemployment," *J. Monet. Econ.*, Apr. 1976, *2*, 185–219.

O. Hart, "On the Optimality of Equilibrium when the Market Structure is Incomplete," *J. Econ. Theory*, Dec. 1975, *11*, 418–43.

B. Holmstrom, "Equilibrium Long-Term Labor Contracts," DP No. 414, Center for Mathematical Studies in Economics and Management Science, Northwestern Univ. 1980.

R. Radner, "Existence of Equilibrium of Plans, Price, and Price Expectations in a Sequence of Markets," *Econometrica*, Mar. 1972, *40*, 283–303.

Technical Change, Returns to Scale, and the Productivity Slowdown

By M. ISHAQ NADIRI AND M. A. SCHANKERMAN*

The recent slowdown in the growth of productivity in the United States has attracted considerable attention. The deceleration has been attributed to many factors, including a slowdown in the growth of capital intensity and the stock of *R&D*, changes in the sectoral composition of output, dramatic rises in oil prices, and declines in the capital utilization rate due to sluggish demand.[1]

In this paper we provide a framework for decomposing changes in total factor productivity (*TFP*) in the presence of economies of scale. The traditional growth accounting framework is a special case of our model. By allowing for economies of scale, we demonstrate formally the positive relationship between growth in productivity and output which is found in the empirical studies by John Kendrick (1973), Nicholas Kaldor, and others.[2] The model is based on an output demand function, a variable (non-*R&D*) cost function which is shifted by disembodied technical change and a stock of *R&D*, and a market-clearing rule which equates output price to average variable cost plus quasi rents to *R&D*. This framework identifies the contribution of demand growth, real factor prices, and the stock of *R&D* to changes in the growth of *TFP*.

As an illustration, we apply the model to American manufacturing for the period 1958–78. The preliminary evidence suggests that the deceleration in demand is a leading factor behind the slowdown in *TFP* growth from 1965–73 to 1973–78. Changes in real factor prices and *R&D* play lesser roles. By contrast, only one-quarter of the decline in *TFP* growth during the earlier period 1958–65 to 1965–73 can be attributed to these factors. These findings contrast sharply with studies which ignore demand shifts by assuming a priori constant returns to scale (for example, Dale Jorgensen and Barbara Fraumeni).

I. The Model

Since we do not impose constant returns to scale, the proper index of conventional *TFP* growth is the "quasi-Divisia" index

$$(1)\quad DTFP \equiv DQ - DF = DQ - \Sigma s_i DX_i$$

where D denotes a rate of growth, Q is output, the X's represent traditional (non-*R&D*) inputs, F is total factor input, and $s_i = P_i X_i / PQ$ is the value share of the ith input.[3]

Let the production function be $Q = G(X, R, T)$ were R and T denote the stock of *R&D* and the (disembodied) technology level. Differentiating with respect to time and assuming cost minimization over all inputs, we obtain

$$(2)\quad DQ = \Sigma[(P_i X_i / Q)/MC]\,DX_i + [(P_r R/Q)/MC]\,DR + DT$$

*New York University and National Bureau of Economic Research. We are indebted to Pierre Mohnen and Richard Nelson for helpful comments, Irene Yew, S. Puglia, and P. Kaplan for very able research assistance and to Elliot Grossman, Jack Faucett Associates, and BEA for providing the data. This research is part of a larger study of productivity and technical change supported by the National Science Foundation, grant PRS-7727048.

[1] For example, see Nadiri, William Nordhaus, and J. Randolf Norsworthy and Laurence Fulco.

[2] For more discussion of this issue, see T. Cripps and R. Tarling, A. Parikh, and W. E. G. Salter.

[3] Charles Hulten shows that if the production function is not linearly homogeneous, the quasi-Divisia index (with value rather than cost shares as weights) must be used in order to preserve path independence of the index. Also see our forthcoming article.

where MC is marginal cost and P_r is the service price, or opportunity cost, of $R\&D$.[4]

The next step is to relate marginal cost to the price of output. We assume that price equals current average variable cost (AVC) plus the unit quasi rents which accrue to past $R\&D$ (see Kenneth Arrow; Ariel Pakes and Schankerman). That is, $P=AVC(1+\theta)$ where θ is the ratio of current quasi rents to the level of AVC. Using the definition of the elasticity along the variable cost function $\eta=MC/AVC$, we obtain $MC=\eta P/(1+\theta)$.[5]

Substituting the expression for MC into (2) we obtain the output growth equation

$$\text{(3)}\qquad DQ=\eta^{-1}(1+\theta)\Sigma s_i DX_i + \eta^{-1}(1+\theta)s_r DR+DT$$

Obtaining $DF=\Sigma s_i DX_i$ from (3) and using (1), the growth of TFP becomes

$$\text{(4)}\qquad DTFP=\frac{(1+\theta-\eta)}{1+\theta}DQ+\frac{\eta}{1+\theta}DT+s_r DR$$

Next we obtain the equilibrium DQ. Assume a *log*-linear per capita demand function. In growth rate form

$$\text{(5)}\qquad DQ=\lambda+\alpha DP+\beta DY+(1-\beta)DN$$

where Y and N are income and population and λ reflects a demand time trend. The pricing rule implies

$$\text{(6)}\qquad DP=DCV-DQ+D(1+\theta)$$

where CV represents total variable cost. The total variable cost function can be written as $CV=H(P_x, Q; R, T_c)$ where T_c is the associated technology level, and both R and T_c shift the variable cost function downward. Factor prices are assumed to be determined outside the model. Differentiating with respect to time, using Shephard's Lemma and the relation $DT_c=-\eta DT$ (see Makoto Ohta), we obtain

$$\text{(7)}\qquad DCV=(1+\theta)\Sigma s_i DP_i+\eta DQ-\Pi DR-\eta DT$$

where $\Pi=P_r R/CV$. Substituting (5)–(7) into (4), we obtain the reduced-form expression for the growth rate of total factor productivity, $DTFP$:

$$\text{(8)}\qquad \begin{aligned} DTFP={}&A[\lambda+\alpha D(1+\theta)]\\ &+A\alpha(1+\theta)\Sigma s_i DP_i\\ &+A\beta DY+A(1-\beta)DN\\ &+s_r[1-A\alpha(1+\theta)]DR\\ &+A\eta(1-\alpha\theta)(1+\theta-\eta)^{-1}DT \end{aligned}$$

where $A=(1+\theta-\eta)[(1+\theta)(1+\alpha(1-\eta))]^{-1}$

Equation (8) decomposes $DTFP$ into four components: 1) factor-price effect, $A\alpha(1+\theta)\Sigma s_i DP_i$; 2) demand effect, $A[\lambda+\beta DY+(1-\beta)DN]$; 3) $R\&D$ effect, $A\alpha D(1+\theta)+s_r[1-A\alpha(1+\theta)]DR$; and 4) disembodied technical change, $A\eta(1-\alpha\theta)(1+\theta-\eta)^{-1}DT$. The underlying model is an equilibrium model in which there is cost minimization over all inputs, the level of $R\&D$ is adjusted until it earns the normal (private) rate of return in the form of quasi rents, and the market clears. Because market clearing is imposed, each component in the decomposition reflects both the direct impact on TFP of the factor in question and its indirect effect via induced changes in the output price. This formalizes Kendrick's suggestion (1973, p. 111) that the relationship between changes in productivity and output growth is "reciprocal" (i.e., DQ leads to $DTFP$ which reduces DP and further raises DQ).

The important parameters in (8) are the price and income elasticities of demand and the cost elasticity of the variable cost function. Note two special cases. First, if demand is completely inelastic ($\alpha=0$), shifts in

[4] An alternative is to restrict cost minimization to the conventional inputs and let $R\&D$ earn a different net rate of return. The service price P_r would then represent the associated gross rate of return (multiplied by an investment goods deflator).

[5] Note that $P=MC$ provided $\eta=1+\theta>1$, i.e., if there are decreasing returns along the variable cost function. This reflects the fact that since $P>AVC$ by a (variable) markup, $P=MC$ can occur only if $MC>AVC$, which implies (local) decreasing returns to conventional inputs.

the cost function due to real factor-price changes ($\Sigma s_i DP_i$) have no effect on output and hence on *TFP*. Second, if marginal-cost pricing prevails ($\eta = 1 + \theta$; see fn. 5), then equation (8) collapses to $DTFP = s_r DR + DT$, which is the standard result when *TFP* is defined only over conventional inputs.[6]

II. An Empirical Application

We now present an empirical illustration of the model. The decomposition requires three parameters, the variable cost elasticity, and the price and income elasticities of product demand. Given these parameters, the procedure is to compute the factor price, demand, and *R&D* effects and then to retrieve the technical change effect as a residual using equation (8).[7] Discrete (Tornquist) approximations to the Divisia indices in (8) are used. Annual data on gross value-added (including energy), capital, labor, and energy for the period 1958–78 were obtained from BEA, Elliott Grossman, and Jack Faucett Associates. Published *NSF* data on *R&D* flows are used to construct a stock, using Kendrick's (1976) benchmark and a depreciation rate of 0.10. Price series on the inputs and outputs were obtained from the same sources and variables such as the rental prices for capital and *R&D* services were generated using familiar formulations. Real factor prices are measured as nominal factor prices deflated by the Consumer Price Index.

TABLE 1—ESTIMATED PRICE, INCOME, AND COST ELASTICITIES

	α	β	η
Manufacturing	−.33	2.16	.76
	(.21)	(.13)	(.05)
Durables	−.20	2.74	.77
	(.21)	(.16)	(.05)
Nondurables	−.56	1.25	.78
	(.18)	(.12)	(.06)

Note: Estimated standard errors are shown in parentheses.

The *log*-linear per capita demand functions are estimated by ordinary least squares with a serial correlation adjustment, using real *GNP* as the income variable. The cost elasticities are obtained from a Cobb-Douglas variable cost function with nonneutral technical change (proxied by time trends). The envelope condition on *R&D* is imposed and estimated together with the variable cost function and the variable-input share equations. The system of equations is estimated by an iterative seemingly unrelated equations procedure with an autocorrelation adjustment. The period covered is 1958–78.

Table 1 summarizes the results. The demand parameters are roughly similar to results reported by Hendrik Houthakker. We also obtained estimates from Data Resources, Inc. which broadly confirm the price elasticities, but place the income elasticity closer to 1.5. A decomposition using this lower value is also performed and the differences will be noted below. The estimated cost elasticities are similar in the three industry groups, are statistically different from unity, and indicate some economies of scale at the industry level. The implied economies of scale may seem somewhat high but they are similar to recent estimates at the industry level by Ernst Berndt and Mohammed Khaled.[8]

[6] In other words, *TFP* growth is equivalent to shifts in the production function provided marginal-cost pricing (and cost minimization) prevails. These shifts represent movements in a technological relationship, but our ability to identify them from market data (*DTFP*) requires behavioral assumptions. Note that long-run constant returns is neither a necessary nor sufficient condition. Of course, the behavioral assumption of marginal cost pricing may only be plausible under nonincreasing returns to scale. See our forthcoming article, and Michael Denny, Melvyn Fuss, and Leonard Waverman.

[7] In the calculations we assume $\Pi = \theta$, or equivalently that the service price of *R&D* equals the current unit quasi rents to *R&D*. This equality reflects the fact that the quasi rents represent the opportunity cost of not "selling" the (rights to the) *R&D*. The assumption $\Pi = \theta$ does not require a normal net rate of return to *R&D*, but our cost minimization assumption does (see fn. 4).

[8] Note two points. First, the null hypothesis $\eta = 1$ is uniformly rejected. The computed χ^2 values are 8.1, 4.0, and 10.0 in manufacturing, durables, and nondurables, respectively, compared to the .05 critical value of 3.84. Constant returns was also rejected in various forms of the translog cost function we estimated. See Berndt and Kahled for the rationale of increasing re-

TABLE 2—DECOMPOSITION OF THE DECELERATION IN TOTAL FACTOR-PRODUCTIVITY GROWTH IN MANUFACTURING INDUSTRIES

Industry	Average $\Delta DTFP$	Factor Prices[a]	Demand[a]	Total Scale[a]	*R&D*[a]	Residual Technical Change[a]
Manufacturing						
1958–65 to 1965–73	−.0157	9.4	13.7	23.1	4.3	72.6
1965–73 to 1973–78	−.0072	33.4	68.3	101.7	17.4	−19.1
Durables						
1958–65 to 1965–73	−.0198	4.4	11.9	16.3	7.2	76.5
1965–73 to 1973–78	−.0095	5.3	63.7	69.0	23.8	7.2
Nondurables						
1958–65 to 1965–73	−.0074	32.5	22.0	54.5	0.9	44.6
1965–73 to 1973–78	−.0063	55.0	44.0	99.0	2.3	−1.3

Note: The parameter estimates for α, β and η used in the decomposition are taken from Table 1.
[a]Shown in percent.

We use these parameter values to decompose the *TFP* growth in total manufacturing, durables, and nondurables for three periods, 1958–65, 1965–73, and 1973–78. In order to focus on the deceleration of *TFP* growth, we present in Table 2 the contributions of factor prices, demand, *R&D*, and residual technical change as a percentage of the change in *TFP* growth between the periods.

The growth in *TFP* declined sharply throughout the entire period. In percentage terms, *DTFP* declined by about 45 percent in manufacturing, 60 percent in durables, and 30 percent in nondurables between each pair of subperiods. The relative contributions of the different factors vary among industry groups and between subperiods. Real factor prices contribute only modestly to the deceleration in *DTFP* from 1958–65 to 1965–73, except in nondurables. Factor prices contribute more to the decline in *DTFP* from 1965–73 to 1973–78. These differences may be partly spurious, and simply reflect the inclusion of the petroleum industry in nondurables.[9] If durables are a good guide, the factor-price effect has been modest.

The slowdown in demand growth is an important factor in the retardation of *DTFP* in all industry groups. About 10–20 percent of the decline in *DTFP* from 1958–65 to 1965–73 is accounted for by the demand effect, but this rises dramatically to more than 50 percent from 1965–73 to 1973–78. Factor prices and demand together (total scale effect) account for most, and in total manufacturing more than all, of the deceleration in *DTFP*. Despite the caveat regarding petroleum and the possibility we have overestimated returns to scale, the evidence points to the scale effect, and mainly demand deceleration, as a major factor behind the decline in *DTFP* from 1965–73 to 1973–78.

The decline in the growth of the *R&D* stock contributes modestly to the slowdown in *DTFP* from 1958–65 to 1965–73, less than 10 percent. The same is true for nondurables from 1965–73 to 1973–78, but in durables and total manufacturing *R&D* plays a very significant role, accounting for nearly a quarter of the retardation in *DTFP*.

The technical change effect is computed residually and therefore captures all contributory factors not accounted for by the model (including measurement error). This residual accounts for the bulk of the decline in *DTFP* from 1958–65 to 1965–73 (about 75 percent in manufacturing and durables and 45 percent in nondurables), but very little from 1965–73 to 1973–78. In fact, the negative contribution in manufacturing and nondurables suggests that residual technical

turns at the industry level. Second, since $\eta < 1$, it follows that the condition for marginal cost pricing ($\eta = 1 + \theta$; note 5) is also rejected.

[9] It would be useful to check this by omitting the petroleum industry and redoing the empirical estimation and the decomposition. We plan to do so in future work.

change accelerated at the same time that *DTFP* declined.[10] The main conclusion is that the deceleration of demand, and to a lesser extent the factor-price and *R&D* effects, dominate the recent slowdown in productivity growth.

III. Concluding Remarks

We propose and illustrate empirically a framework for decomposing changes in *TFP* in the presence of economies of scale. The traditional growth accounting framework with constant returns is a special case of our model. The empirical results suggest that the deceleration of demand is a leading factor in the recent decline of *TFP* growth in American manufacturing. Our analysis underscores the importance of understanding the forces behind the slowdown in demand, but we do not address this question here.

There are several ways to extend the analysis: 1) explore more fully the link between the quasi rents and the service price of *R&D*; 2) endogenize the determination of factor prices; and 3) relax the assumption that *R&D* earns the normal private return, estimate the realized rate of return in the model, and use it to evaluate the impact of *R&D* on *TFP* growth.

[10] These qualitative conclusions are unaffected if we substitute the smaller income elasticities of demand for durables and manufacturing ($\simeq$1.5) provided by DRI. There is a negligible change in the residual technical change effect from 1958-65 to 1965-73; it becomes zero in manufacturing and 15 percent in durables from 1965-73 to 1973-78. The sharp reversal of the importance of the residual still holds.

REFERENCES

K. J. Arrow, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Universities–Nat. Bur. Econ. Res. conference series, Princeton 1962.

E. R. Berndt and M. S. Khaled, "Parametric Productivity Measurement and Choice Among Flexible Functional Forms," *J. Polit. Econ.*, Dec. 1979, *87*, 1220–45.

T. Cripps and R. Tarling, "Growth in Advanced Capitalist Economies 1950–70," Dept. of Applied Economics Occasional Paper 40, Cambridge Univ. 1973.

M. Denny, M. Fuss, and L. Waverman, "The Measurement and Interpretation of Total Factor Productivity in Regulated Industries, With an Application to Canadian Telecommunications," in Tom Cowing and Rodney Stevenson, eds., *Productivity Measurement in Regulated Industries*, New York forthcoming.

H. Houthakker, "Growth and Inflation: Analysis by Industry," disc. paper no. 704, Harvard Inst. Econ. Res., May 1979.

C. Hulten, "Divisia Index Numbers," *Econometrica*, Nov. 1973, *41*, 1017–26.

D. Jorgensen and B. Fraumeni, "Relative Prices and Technical Change," in E. R. Berndt and B. C. Field, eds., *The Economics of Substitution in Production*, Cambridge, Mass. forthcoming.

Nicholas Kaldor, *Causes of the Slow Rate of Economic Growth of the United Kingdom*, Cambridge 1966.

John Kendrick, *Postwar Productivity Trends in the United States 1948–1969*, New York 1973.

______, *The Formation and Stocks of Total Capital*, New York 1976.

M. I. Nadiri, "Sectoral Productivity Slowdown," *Amer. Econ. Rev. Proc.*, May 1980, *70*, 349–52.

______ **and M. Schankerman,** "The Structure of Production, Technological Change, and the rate of Growth of Total Factor Productivity in the U.S. Bell System," in Tom Cowing and Rodney Stevenson, eds., *Productivity Measurement in Regulated Industries*, New York forthcoming.

W. D. Nordhaus, "The Recent Productivity Slowdown," *Brookings Papers*, Washington 1972, *73*, 493–546.

J. Norsworthy and L. Fulco, "Productivity and Costs in the Private Economy," *Mon. Labor Rev.*, May 1976, *99*, 3–11.

M. Ohta, "A Note on the Duality Between Production and Cost Functions: Rate of

Returns to Scale and Rate of Technical Progress," *Econ. Stud. Quart.*, Dec. 1974, *25*, 63–65.

A. Pakes and M. Schankerman, "An Exploration into the Determinants of Research Intensity," Nat. Bur. Econ. Res., work. paper no. 438, Jan 1980.

A. Parikh, "Differences in Growth Rates and Kaldor's Laws," *Economica*, Feb. 1978, *45*, 83–92.

W. E. G. Salter, *Productivity and Technical Change*, 2d ed., Cambridge 1966.

Public Regulations and the Slowdown in Productivity Growth

By GREGORY B. CHRISTAINSEN AND ROBERT H. HAVEMAN*

Since 1965, indices of labor productivity have had a disappointing and largely unexplained performance. Not only is the rate of productivity growth over the post-1965 period lower than in preceding postwar years, but its upward trend has been broken at least twice. Since 1978, productivity growth has been effectively zero. If the trend of labor productivity from 1946–65 had continued until 1980, the current index would be about 15 percent above its actual level. Table 1 summarizes the postwar behavior of four alternative measures of productivity.

While productivity growth has slowed in nearly all sectors, there is a large variance in the distribution of post-1965 sectoral productivity growth rates. The most dramatic slowdowns have been recorded for the mining, utilities, and construction sectors. The manufacturing sector has experienced a much milder slowdown, and since 1967 its productivity index has risen over 12 percentage points more than that for the entire nonfarm sector.

Many phenomena have contributed to poor productivity performance. They range from subtle changes in worker motivation to the propensity to innovate in both products and processes to exogenous shocks to the production process (due, for example, to unexpected energy price changes) to alterations in output mix, the demographic characteristics of the labor force, or the ratio of labor to capital to the nature and intensity of regulatory policy. Not only are these effects numerous, but they interact in complex and dynamic ways. Numerous assertions have been made regarding the contribution of each of these factors, and studies seeking to identify their relative contributions have been undertaken. At present, the contribution of public regulations to both the productivity slowdown and to poor economic performance generally is both widely debated and little understood. It is this relationship that is the focus of this paper. In Sections I and II, the direct and indirect ways in which regulations can adversely affect productivity are distinguished, and the existing studies of this relationship are described. Sections III and IV describe our attempts to model and estimate the contribution of regulation to the slowdown. Some preliminary results are presented.

I. Public Regulations as a Source of the Productivity Slowdown

By definition, public regulations are interventions into market processes. Because of them, the utility and profit-maximizing decisions of individual decision makers are altered. In a smoothly functioning market economy (without externalities), such interventions ensure deviation from the private sector production frontier. Holding output composition constant, this deviation means that additional inputs are required to produce any given level of output. Under these conditions, increases in the intensity of public regulations will be associated with larger deviations from the private output frontier, and equivalently, reduced rates of growth of output per unit of input—productivity. In a dynamic setting, increased regulatory intensity, through its alteration of private optimizing decisions, is likely to induce reductions in the measured rate of productivity growth.

The channels by which public regulations are likely to affect either the output numera-

*Assistant professor of economics, Colby College, and professor of economics, University of Wisconsin-Madison, respectively. Helpful comments on an earlier draft by John Bishop, Laurits Christensen, Sheldon Danziger, Donald Nichols, Eugene Smolensky, and Barbara Wolfe are gratefully acknowledged.

TABLE 1–POSTWAR ANNUAL PRODUCTIVITY GROWTH RATES IN THE UNITED STATES, VARIOUS MEASURES OF PRODUCTIVITY
(Shown in Percent per Year)

	Output per Person-Hour, Private Sector	Output per Person-Hour, Nonfarm Private Sector	Nonresidential Business Income per Person Employed	Total Factor Productivity in Domestic Private Business
1947-66	3.44	2.83	2.9	2.9[a]
1966-73	2.15	1.87	1.3	1.4[b]
1973-78	1.15	1.02	–.1	–
1979	–.9	–1.2	–	–

Source: Figures for output per person-hour, private sector and output per person-hour, nonfarm private sector were taken from Jerome Mark, p. 486. Figures for nonresidential business income per person employed were taken from Denison (1979c), p. 21. Figures for total factor productivity in domestic private business were taken from Kendrick, p. 511.

[a] For years 1948-66.

[b] For years 1966-76.

tor or input denominator of productivity indices are complex. Each channel involves some aspect of policy-induced business behavior entailing a reduction in the ratio of output to input. To illustrate these channels, we will deal with environmental regulations; analogous channels of impact exist for other forms of regulation.

By their nature, environmental regulations require investments to reduce residual flows. To the extent that these investments compete with standard plant and equipment investments, the ratio of labor to conventional capital will be increased. Moreover, because these regulations are typically based on engineering standards, the activities which they generate tend to be excessively capital intensive. Because these regulations fall especially heavily on new pollution sources, incentive is given for uneconomic retention of existing—and lower productivity—plant and equipment. These regulations have also tended to be more heavily imposed on sectors with high postwar rates of productivity growth (for example, utilities), and in low pollution regions attractive for plant location. And, because pollution control equipment requires manpower to operate it, employment levels rise with no addition to marketable output. Finally, complying with these regulations requires the information-gathering, administrative, and legal activities which require inputs yielding no saleable output. Meeting these requirements may also require time—causing delay in expansion and modernization plans and the stretching-out of construction periods.

II. Public regulations and the Productivity Slowdown: Some Estimates

No comprehensive study of the effect of public regulations on the slowdown in productivity growth has been undertaken. A few studies of the contribution of environmental and health/safety regulations have been made, however.

The most influential of these is that of Edward Denison (1979a, b, c) who uses his growth accounting framework to derive an estimate of the contribution of these regulations to the retardation of growth in his productivity measure—final output valued at factor cost per unit of labor, capital, and land inputs. Denison's index suggests that the average annual impact of post-1967 environmental regulations on the rate of productivity growth was .05 percentage points from 1967–69, .1 percentage points from 1969–73, .22 percentage points from 1973–75, and .08 percentage points from 1975–78. Robert Crandall has also studied the environmental regulation-productivity interaction, using both cross-section and time-series regression approaches. While the results from these estimates vary substantially, he finds that the index of manufacturing in 1976 is about 1.5 percent below what it would be in the absence of mandated pollution control expenditures, and that those manufacturing sectors heavily impacted by environmental regulations showed a greater slowdown in productivity growth after 1970 than manufacturing as a whole. Finally, Robin Siegel has attempted to account for the slowdown in the private nonfarm labor productivity trend by regressing this quarterly time-series variable on variables designed to account for changes in output due to the business cycle, output and labor force composition, relative energy prices, pollution control expenditures, and capital investment, among other potential factors. Pollution control expenditures were estimated to have caused a 0.5 percentage-point reduction in the rate of productivity growth from 1965–73, but no significant effect after 1975.

These studies have focused on pollution control (as opposed to the full set of public) regulations. They have not been based on a rigorous theoretical or estimation framework, and omitted variable and other data and statistical problems plague the estimates. Elsewhere, we have critiqued these and other studies and the estimates which they have yielded, concluding that between 8–12 percent of the post–1973 slowdown in the growth rate of labor productivity is attributable to environmental regulations (see our paper with Frank Gollop).

III. The Productivity Impact of Regulation: An Empirical Framework

To provide a preliminary evaluation of the contribution of public regulations to the slowdown in productivity growth, we employ a simple time-series regression model for the *U.S.* manufacturing sector. We assume that there is a differentiable aggregate production function underlying economic activity in the manufacturing sector which relates the flow of output (Q) to the flow of total factor input (TFI). The function shifts over time (T) and also in response to what we refer to as "regulatory intensity" (R).

Assuming constant returns to scale, a simple first-order form is

$$(1) \qquad Q = A(TFI)\cdot e^{\alpha R+\beta T}$$

where A, α, and β are parameters. Taking the natural logarithm of both sides of (1):

$$(2) \qquad ln\,Q = ln\,A + ln(TFI) + \alpha R + \beta T$$

If $ln(TFI)$ is subtracted from both sides of (2), an equation for the level of total factor productivity (TFP) is obtained:

$$(3) \qquad ln(TFP) = ln\,A + \alpha R + \beta T$$

That is, to the extent that economic activity follows the hypothesized production function, the level of total factor productivity is a function of a constant, regulatory intensity, and time.

We assume that production in the *U.S.* manufacturing sector can be approximated by (1) except during periods in which the sector is "shocked" by business-cycle effects. Accordingly, in addition to an additive disturbance term (v), we add two terms to (3) designed to capture cyclical effects on total factor productivity. Following William Nordhaus, who has justified this procedure in a more rigorous setting, the additional variables are current and lagged values of $ln(Q/Q^*)$, where Q is actual output and Q^* is a measure of the level of output which would have been produced in the absence of cyclical influences.[1]

Thus, our equation for the level of total factor productivity is

$$(4) \qquad ln(TFP) = ln\,A + \alpha R + \beta T + \gamma\,ln\left(\frac{Q}{Q^*}\right) + \delta\,ln\left(\frac{Q}{Q^*}\right)_{-1} + v$$

[1] Assume that the actual level of output (Q) depends on the level of demand which, in turn, depends on a constant, the price level of sector output relative to the general price level, the deviation of the actual from the "natural" rate of unemployment ($U-U^*$), and "natural" real GNP (GNP^*). The Q^* is estimated by regressing Q on its determinants and imputing values of Q assuming $U=U^*$. (The U^* and GNP^* are from Robert Gordon.)

where γ and δ are parameters, and R enters the equation with an as yet unspecified lag distribution. As is well known, *TFP* differs from *labor* productivity by a factor reflecting the influence of the ratio of nonlabor to labor inputs (K/L).

We have estimated (4) for the *U.S.* manufacturing sector from 1958–77 using unpublished annual data on the quantities and proportions of total cost accounted for by labor, capital, energy, and materials, and price and quantity data pertaining to output.[2] In order to reduce the presence of multicollinearity, these inputs were combined into a measure of *TFI* by using their respective shares in total cost as weights. Because of this comprehensive set of inputs, the effect of some factors often assigned responsibility for the productivity slowdown (for example, the energy crisis) is filtered out of the *TFI* measure.

"Regulatory intensity" is a difficult concept to define, let alone quantify. As noted, our definition of this concept is based on the view that public regulatory agencies distort optimizing private sector decisions which would, *ceteris paribus*, maximize the measured rate of productivity growth. We have constructed three alternative indices of this variable for the postwar period. The first is based on an estimate of the cumulative number of "major" pieces of regulatory legislation in effect during any of the years in question (R_1).[3] The second and third indices are based on the volume of real federal expenditures on regulatory activities for the years in question (R_2) and the number of full-time federal personnel engaged in regulatory activities (R_3).[4] For our measures, that portion of each agency's activities devoted to the manufacturing sector was the average of the judgments of several recognized students of regulation.[5] Though crude proxies for regulatory intensity, we believe these indices provide a reasonable characterization of postwar trends in the regulation of the manufacturing sector; indeed, the only characterization available without a major research effort.

Each of the R indices imply only a gradual increase in regulatory intensity until the mid-1960s. Then, all three measures accelerate, with R_2 increasing at a more rapid rate than R_3 which in turn shows a greater acceleration than R_1. All of the measures show a further acceleration during the 1970's, though the acceleration is again least pronounced in the case of R_1. Setting each index equal to 100 in 1947, R_1 attains a level of 402.88 in 1977, while R_2 and R_3 read 1003.77 and 668.03 respectively. While there are exceptions, the indices generally imply a monotonic increase in regulatory intensity during the 1947–77 period.

Alternative estimates of equation (4) were obtained using R_1, R_2, and R_3, with lag specifications chosen on the basis of the Bayesian estimation criterion proposed by John Geweke and Richard Meese. A simple one-year lag was chosen for R_2 and R_3; two-years for R_1. So lagged, the simple correlation coefficients among the alternative measures are: .85 (R_1, R_2), .86 (R_1, R_3), and .94 (R_2, R_3). Pseudo-generalized least squares estimates of the equation were made by using Takeshi Amemiya's procedure for prefiltering the data.

IV. The Productivity Impact of Regulation: Preliminary Results

Combining our regression estimates with estimates of the impact of K/L accounted for by differences in the growth rates of *TFP* and labor productivity, we obtain the results in Table 2.

[2] We wish to thank J. R. Norsworthy and Michael Harper of the U.S. Bureau of Labor Statistics for these data.

[3] This series was calculated from data presented in Center for the Study of American Business.

[4] R_2 and R_3 were estimated from agency data published in the *Budget of the United States Government*. For large, diverse agencies such as the Environmental Protection Agency, data on regulatory functions are separable from other agency functions. For smaller regulatory agencies, we have used expenditure and staffing data for the agency as a whole.

[5] Each individual was asked to estimate the percentage of each agency's activities which are devoted to the manufacturing sector, and how this percentage had changed over time. In each case, the highest and lowest estimates were discarded, and the mean of the remaining estimates was used in constructing these indices.

TABLE 2—CONTRIBUTIONS TO THE RATE OF GROWTH OF LABOR PRODUCTIVITY IN *U.S.* MANUFACTURING, 1958–77: PRELIMINARY RESULTS

	Contribution during:		
Source	1958-65	1965-73	1973-77
R	0 to $-.1$	$-.1$ to $-.3$	$-.2$ to $-.3$
T	.9 to 1.0	.9 to 1.0	.9 to 1.0
Q/Q^*	0 to .1	0	0 to $-.1$
Unexplained	.4 to .5	$-.1$ to $-.2$	$-.3$ to $-.4$
Average Growth Rate of Total Factor Productivity	1.4	.6	.3
K/L	1.6	1.9	1.4
Average Growth Rate of Labor Productivity	3.0	2.5	1.7

These numbers are derived by simply taking the parameter estimates for equation (4) and then multiplying them by the average annual changes in the associated variables. In the case of R, the estimated regression coefficients for α are .011 (R_1), .005 (R_2), and .006 (R_3). Lagged appropriately, the average annual changes in R for the three periods of Table 2 are 6.66, 13.48, and 20.44 (R_1), 6.09, 65.90, and 61.32 (R_2), and 5.05, 33.61, and 49.80 (R_3). The average percentage point contributions of regulation to the rate of growth of total factor productivity are then calculated to be $-.073$, $-.148$, and $-.224$ (R_1), $-.030$, $-.330$, and $-.301$ (R_2), and $-.030$, $-.202$, and $-.299$ (R_3).

Neither R_1 nor R_3 was statistically significant at either a .01 or a .05 level, but both were significant at a .10 level. The estimated coefficient for R_2 was significant at the .05 level. In all cases, neither the estimated coefficient on the lagged cyclical variable nor an interaction term for R and T were significant at the .10 level. The same was true of an interaction term for R and K/L in an equation for the level of labor productivity. All other estimated coefficients were significant at the .05 level.

The ranges indicated in Table 2 thus stem from the alternative measures of R. Of the alternatives, R_2 (which shows the greatest acceleration in regulatory intensity over time) implies the most negative impact on the rate of productivity growth. It also implies the greatest rate of "technical change" and the smallest average cyclical impact. These conclusions are reversed for R_1, with those for R_3 being intermediate to the other two.

V. Summary and Caveats

These results suggest that federal regulations are responsible for from 12 to 21 percent of the slowdown in the growth of labor productivity in *U.S.* manufacturing during 1973–77 as compared to 1958–65.[6] They are consistent with previous research noted in section II. Reductions in the ratio of nonlabor to labor inputs (K/L) are responsible for about 15 percent of the slowdown. The contribution of the average cyclical impact could fall anywhere in the 0–15 percent range. The unexplained portion of the slowdown in the rate of productivity growth —often attributed to changes in labor force composition, *R&D* expenditures, or sectoral output shifts—remains substantial.

These results on the impact of regulation are, in certain important respects, sensitive to the manner in which the model is specified. For example, with alternative lag specifications, estimated coefficients for R may be insignificant. Also, if a separate trend variable for each of the three periods in Table 2 is entered into the model, or if time enters in second-order form, multicollinearity among the explanatory variables causes the coefficients for both regulatory intensity and time to be insignificant. Moreover, the R variables may be capturing other exogenous forces inducing contemporaneous productivity growth reductions, in

[6] This conclusion is derived by taking the difference between the percentage point contributions of each regulatory intensity variable during 1958–65 and 1973–77, and dividing this difference by the difference between the rates of growth of labor productivity during the two periods. These values are all shown in Table 2.

which case improved specifications may reduce the estimated R impacts. While we believe that our 12–21 percent estimated contribution of regulatory intensity to the slowdown in the growth of labor productivity will prove to be robust with respect to improved data and more sophisticated models,[7] we recognize the uncertainties surrounding this estimate caused by less-than-ideal data and the possible recent impact on productivity growth of many other factors —factors which may be difficult to measure or capture in any simple model. It should be noted, however, that the procedure employed is more robust with respect to assumptions than those used in widely quoted studies which estimate the response of investment spending to taxation, private savings to Social Security wealth, or productivity growth to *R&D* spending.

Finally, our study focuses on the contribution of public regulations to *measured* productivity. Such regulations are typically undertaken in the belief that they will yield contributions to economic welfare not fully reflected in measured output (for example, improved health and safety; an improved environment). If such gains are forthcoming, growth in "true" economic productivity would exceed its measured counterpart. Our results have little implication for the contribution of public regulations to true productivity growth.

[7] The impact of R was also estimated using 1947–71 data on prices, quantities, and proportions of total cost accounted for by labor, capital energy, and materials compiled by Ernst Berndt and David Wood. This series was extended to 1977 by applying estimated percentage changes in each variable indicated by the *BLS* data, and normalizing cost shares. Because the variable definitions in the two data sets are not identical this exercise, taken by itself, would be of dubious value. This estimation implied an impact on *BLS*-defined labor productivity in the 12–25 percent range, however.

REFERENCES

T. Amemiya, "Generalized Least-Squares with an Estimated Autocovariance Martix," *Econometrica,* July 1973, *41,* 723–32.

E. R. Berndt and D. O. Wood, "Technology, Prices, and the Derived Demand for Energy," *Rev. Econ. Statist.,* Aug. 1975, *62,* 259–68.

G. Christainsen, F. Gollop, and R. Haveman, "Environmental and Health-Safety Regulations, Productivity Growth, and Economic Performance: An Assessment," Joint Economic Committee, U.S. Congress, 1980.

R. Crandall, "Pollution Controls and Productivity Growth in Basic Industries," in Thomas G. Cowing and Rodney Stevenson, eds., *Productivity Measurements in Regulated Industries,* New York forthcoming.

Edward F. Denison, (1979a) "Pollution Abatement Programs: Estimates of Their Effect Upon Output Per Unit of Input, 1975–1978," *Surv. Curr. Bus., Part I,* Aug. 1979, *59,* 58–59.

______, (1979b) "Explanations of Declining Productivity Growth," *Surv. Curr. Bus., Part II,* Aug. 1979, *59,* 1–24.

______, (1979c) *Accounting for Slower Economic Growth,* Washington 1979.

J. Geweke and R. Meese, "Estimating Regression Models of Finite But Unknown Order," SSRI Paper no. 7925, Univ. Wisconsin-Madison, 1979.

Robert J. Gordon, *Macroeconomics,* Boston 1978.

J. Kendrick, Testimony before the Congressional Joint Economic Committee, in *Special Study on Economic Change: Hearings before the Joint Economic Committee, Congress of the United States,* Part 2, Washington 1978, 616–36.

J. Mark, Testimony before the Congressional Joint Economic Committee, in *Special Study on Economic Change: Hearings before the Joint Economic Committee, Congress of the United States,* Part 2, Washington 1978, 476–86.

W. O. Nordhaus, "The Recent Productivity Slowdown," *Brookings Papers,* Washington 1972, *3,* 473–546.

R. Siegel, "Why Has Productivity Slowed Down?," *Data Resources Rev.,* Mar. 1979, *1,* 1.59–1.65.

Center for the Study of American Business, *Directory of Federal Agencies,* Formal Publication No. 31, St. Louis 1980.

The Productivity Growth Slowdown and Capital Accumulation

By MARTIN NEIL BAILY*

The purpose of this paper is to review the role of capital in the slowdown and to assess the key policy issue: Should the rate of capital accumulation be raised as a response to the slowdown?

I. What Has Happened?

According to the Bureau of Labor Statistics (*BLS*), labor productivity grew at an average annual rate of 2.57 percent a year from 1948–68 in the nonfarm business sector of the economy (here and elsewhere percentage changes are expressed as 100 times the change in the natural logarithm). This rate of growth fell to 1.64 percent from 1968–73. It fell again to 0.92 percent from 1973–77 and the level of labor productivity has actually declined slightly from 1977–79, despite a fairly rapid growth of output (3.76 percent a year).

Two major growth accounting studies (by Edward Denison, and by Barbara Fraumeni and Dale Jorgenson) have examined the period from 1948–76 in order to allocate output growth determinants to capital, labor, and total factor productivity (*TFP*). Despite important differences between the two studies, the conclusions are rather similar. I fit a cyclical correction factor and various time trends to the *log* of *TFP* computed in each study. Both showed two distinct breaks in the *TFP* trend. The *TFP* growth declined by about one percentage point in the 1967–69 period and by about an additional one and one-half points after 1973.

I had thought it likely that a gradual slowdown in *TFP* growth combined with a sharp break after 1973 would show up as the pattern, consistent with, say, a slow weakening of the pace of new innovation plus a cyclical or oil-related shock after 1973. But for both sets of figures, a quadratic time trend with a single break performed noticeably worse than two breaks. The accounting studies, therefore, suggest that the slowdown in labor productivity growth through 1976 was associated with a two-step decline in *TFP* growth. The *BLS* numbers for 1977–79 raise the possibility of yet a third slowdown. Whether or not this also is a *TFP* slowdown remains to be seen.

The obvious candidate to explain a slowdown in labor productivity growth is capital. But since the two accounting studies show sharp *TFP* slowdowns, it follows that neither will show a decline in capital accumulation to have been a major cause of the decline in labor productivity growth. In fact, in Jorgenson's calculation the capital-labor ratio, which grew at 2.73 percent per annum from 1948–76, grew at 2.55 percent per annum from 1973–76. Comparing the 1948–68 period with 1968–76, Jorgenson finds that a decline in *TFP* growth accounts for virtually all of the decline in labor productivity growth. Denison's findings are similar.

Another aspect of capital and the slowdown concerns the pre-tax return on capital. If *TFP* growth has slowed, then a simple growth model predicts that the rate of profit will decline for a given rate of capital accumulation. If *TFP* growth has maintained a constant pace, but the rate of capital accumulation has slowed, then the rate of profit should rise. The calculations by Martin Feldstein and Lawrence Summers for the nonfinancial corporate business sector 1948–76 show no dramatic trends in the return to capital, but a dummy variable for 1970–76 shows a weakly significant drop in the profit rate of one to one and one-half percentage points.

*Senior Fellow, The Brookings Institution. This research has been supported by the Sloan Foundation and the U.S. Department of Labor. Discussions with and the written work of Ernst Berndt, Barry Bosworth, Edward Denison, Dale Jorgenson, William Nordhaus, and Robert Solow have been helpful. Lorin Kusmin and Lesley Kalmin provided excellent assistance.

It is possible to make a case that capital does have a more important role in the slowdown than suggested above (see Section III). But for the present, accept as a tentative hypothesis that the rate of technical change has in fact declined, and now consider whether or not this should prompt policy action to stimulate more rapid capital accumulation as an offset.

II. A Productivity Slowdown and Optimal Capital Accumulation

Consider the implication of an optimal growth model following a slowdown in the rate of technical change. The structure of the model and the method of analysis used will follow the treatment in Kenneth Arrow and Mordecai Kurz. The objective function is

$$\int_0^{\infty} e^{-\rho t}P(t)U\{c(t)\}dt \tag{1}$$

where P is population, U is the utility function, assumed to have a constant elasticity of marginal utility σ, and c is per capita consumption. The rate of time discount of utility is ρ. The objective function is maximized subject to the constraint that output be divided between investment and consumption.

The optimal path can be elaborated more easily by defining k and λ as follows:

$$k=\frac{K}{Le^{\gamma t}} \qquad \lambda=U'(c)\left[\frac{Le^{\gamma t}}{P}\right]^{\sigma} \tag{2}$$

where K is the capital stock, γ is the rate of Harrod-neutral technical progress, L is the labor force, and it is assumed that L and P grow at the same rate n. The nature of the optimal trajectory is illustrated in Figure 1; refer to the solid lines. The normal case is when the economy is on the segment AB, i.e., it is raising its capital to effective-labor ratio k and converging to the steady-state optimum at B, with λ^* and k^*. The optimal turnpike is achieved when the MP of capital exceeds the time discount rate, i.e., where $f'(k)=\rho+\sigma\gamma$.

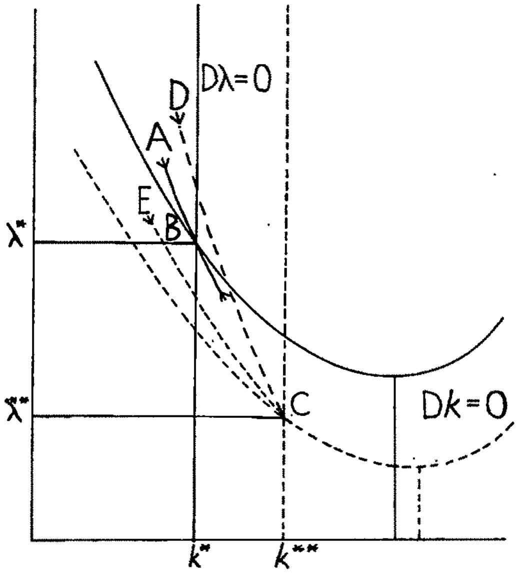


FIGURE 1. OPTIMAL ACCUMULATION TRAJECTORIES

The impact of a decline in the rate of technical progress is shown by the dotted lines in Figure 1. The $D\lambda=0$ locus has been displaced to the right by the decline in γ. Thus the steady-state optimal growth turnpike requires a higher value of k. The $Dk=0$ locus has been shifted downwards and its minimum point shifted to the right. The new intersection with the $D\lambda=0$ locus occurs at C. If the economy was initially somewhere on the AB trajectory, then optimality requires moving to a new trajectory through C. Two possibles are shown as DC and EC. In the case DC, the shift requires an increase in λ, starting from a given initial k. The opposite case is implied by the trajectory EC. If we compare the old and new turnpikes, we see that, along the post-slowdown turnpike, λ^{**} is less than λ^* but k^{**} is greater than k^*. This means that the long-run optimal savings propensity may also rise or fall as a result of the decline in γ. The most important unknowns in this long-run outcome are the values of σ and ρ. If preferences are very "risk averse" (σ large), the savings rate increases. If ρ is large, the rate declines.

Thus the conclusion from optimal growth analysis is that if an economy is on an optimal trajectory and technical progress diminishes, there is no very strong case either way for a change in the fraction of output accumulated in the short run or the long run. This conclusion *may* be dependent on the assumption of Harrod neutrality. It is not certain it would be, because when the elasticity of substitution is close to unity there is no effective difference among the Hicks, Harrod, and capital-augmenting technical change alternatives. And as Ernst Berndt and Jorgenson have pointed out, a large body of research into production functions supports the assumption of a long-run unitary elasticity.

If an economy is not on an optimal path to begin with, then presumably the case for altering the rate of accumulation is not clearly affected by a decline in technical progress.

III. The Role of Capital Once Again

In many discussions of the slowdown a much larger role is given to capital than the one attributed by Denison and Jorgenson. But the standard data through 1976 simply do not show this. Since 1976 the rapid growth of employment has slowed the growth of the capital-labor ratio somewhat, which bears on the third step of the slowdown. But if one looks at the whole period from 1948 to the present, the match between the growth of capital intensity and the growth of labor productivity just is not very close. (The careful study by Randall Norsworthy et al. is probably about as far as one can go with capital using conventional methods.)

The trouble with this conclusion, though, is that the story from the front lines sounds different. Businessmen and the business press suggest that much of the capital in place is obsolete or ill-suited to current conditions and must be replaced. It is argued that there is a great need (however defined) for new capital investment, and that profit streams are not generating the necessary funds. Now of course we cannot discard hard evidence for anecdote and of course we must recognize that a desire for tax cuts may color opinions. But there is surely a case for questioning the way in which capital is measured.

The trouble with the standard data is that there is no direct evidence on the scrapping of old capital or the extent to which old capital may embody a technology not directly comparable to that of new capital. The principal source of information on scrapping is IRS Bulletin F, published in 1942 and based upon pre-World War II evidence. Some postwar information has been used to modify Bulletin F (equipment lives are reduced by 15 percent for example), but no adjustment is made to the capital stock series depending on whether new investment adds to or modifies existing capital or instead replaces basically obsolete plant. This is not a complaint about methodology. The necessary information just is not available.

The Commerce Department may have understated effective capital stock growth from the mid-1950's to the mid-1960's and overstated capital growth from the late-1960's to the present for two major reasons: (i) an echo effect from the depression and World War II; (ii) major structural changes occurring in the late 1960's and the 1970's.

In 1946, the *U.S.* economy had just emerged from sixteen years in which investment had been held down by depression and war-time controls. In the immediate postwar years, a high level of capital spending gradually rebuilt the capital stock. Investment then stayed on a plateau from 1955–64, but effective capital stock growth probably exceeded measured capital stock growth, because the new plant and equipment supplemented and modified the capital already in place. Starting in the late 1960's, many industries found that the basic design and production method embodied in their existing plant and equipment had become obsolete. Adding incremental new capital would yield only a small productivity advantage. Major new investment was necessary if firms were to make use of continuing technological developments. This situation was compounded by many structural changes in the 1970's, such as changing comparative advantage, environmental and safety regulations, the rise in energy prices and the unwillingness of consumers to buy V-8 engines. The investment that has taken place has not resulted in capital deepening because of a high rate of scrapping of old capital or because the old capital is of limited economic value.

It also follows that if the capital stock has not been correctly measu: ed, then neither has the rate of profit. In particular, the rate of profit on the effective capital stock in the late 1970's may have been much higher than the measured rate.

IV. Market Valuation of the Capital Stock

Is there any hard evidence to support the ideas just presented? Such evidence may be provided by the market valuation of the capital stock. There has been extensive discussion in the literature of three related puzzles: (a) Given the estimated size of the capital stock, why was the market valuation of the corporate sector in the late 1970's so low? (b) Given the low market valuation, why did firms continue to invest in new plant and equipment? (c) Given the estimated size of the capital stock, why was productivity so low? All three puzzles disappear or become less serious if in fact the standard estimates of the size of the capital stock substantially exceed the actual or effective capital stock.

Several explanations have been offered for the low level of the stock market. These include increased corporate taxes, higher interest rates used to discount future earnings, and pessimism about the growth rate of earnings. There is surely merit in these suggestions, but whether true or not, they do not solve the puzzles as stated. The key concept is q (the ratio of the market value of corporations to the replacement cost of their net assets) developed by James Tobin. According to q theory, corporations should accumulate capital if $q>1$ and decumulate if $q<1$ such that q remains close to unity. Swings of mood or expectation in the stock market can very well cause short-run deviations of q from unity, but any systematic forces raising or lowering the market valuation of the corporate sector should stimulate a change in the rate of capital formation to restore equilibrium. In fact, however, q has been below unity since 1968, according to George von Furstenberg. It has declined steadily since the late 1960's and was less than 0.6 in 1978. In the same year, corporations spent $163 billion on new equipment and structures. The discussion of this puzzle has largely been in terms of the possible deficiencies of q theory. However, given the assumptions, the theory is impeccable and the assumptions are basically the same as those of other neoclassical production and investment models. In equilibrium, the size of the capital stock is such that the price of a unit of capital p_K is just equal to the present discounted value of the stream of quasi rents $[r(t)]$.

$$p_K=\int e^{-\rho t}r(t)dt \tag{3}$$

where ρ is the discount rate. If K is the number of units of capital in existence, then multiply both sides of (3) by K to give

$$p_K K=V \tag{4}$$

where V is now the market valuation of the capital stock. The extent to which the standard data for $p_K K$ differ systematically from V (i.e., the extent to which q differs from unity) then provides an estimate of the extent to which the standard figure for K differs from the effective capital stock.

V. Growth in the Nonfinancial Corporate Sector

If output is produced by capital and labor, then the standard growth accounting relation is the following:

$$\Delta \ln Q=\alpha\Delta \ln L+(1-\alpha)\Delta \ln K+v \tag{5}$$

where Q, L, and K are output, labor and capital, and α is the share of labor in total cost. This equation then defines v the growth rate of *TFP* from $t-1$ to t. A decomposition of labor productivity growth into *TFP* and capital contributions in the nonfinancial corporate business sector is shown in Table 1, using the standard capital stock data and then a "market-value" adjusted capital series

TABLE 1—DECOMPOSITION OF LABOR PRODUCTIVITY GROWTH

Period	Labor Productivity Growth	Standard Data		Market Valuation	
		TFP Contribution	Capital Contribution	*TFP* Contribution	Capital Contribution
1959-69	2.76	2.73	0.03	2.63	0.13
1969-78	1.43	1.31	0.12	2.36	−0.93

Notes: Output, Labor input, and net stock of equipment and structures for the nonfinancial corporate business sector from the DRI data base. Output and capital are in constant dollars. Market valuation of the net stock is computed as q times the net stock above. (q is from von Furstenberg.) α_t is the labor share in total cost averaged over $t-1$ and t. The capital share is $1-\alpha_t-0.1$.

(defined as the standard series times q). An awkward question in growth accounting concerns the extent to which different kinds of capital assets actually contribute to *measured* physical productivity growth. Some assets are held because of customer convenience or because they yield capital gains. Only a tiny fraction of the total holdings of land are actually involved in production in any given year. I plan to explore this question further in future work, but for the present, K is restricted to include only equipment and structures and its share weight in equation (5) is therefore reduced.

There are substantial reasons to view any growth accounting with caution, and the nonfinancial corporate sector is not ideal for productivity analysis. Further, the general pattern of q was well known before this accounting exercise was carried out. Nevertheless, it is of interest that, unlike the standard capital measure, the q-adjusted or "market value" measure attributes most of the slowdown in labor productivity growth to capital. That is probably too much of a good thing. It is likely that 1978 was a disequilibrium year so that the q-adjusted capital measure is understating the effective capital stock. But it is suggestive all the same.

VI. Conclusions

The standard story from the standard data says that labor productivity growth has declined because *TFP* growth has declined (except perhaps for 1976–79.) If this is true, there may be a case for increasing the rate of capital accumulation if it was already too low, but there is no clear case for a big push to raise capital growth to *offset* the *TFP* decline.

The results from using the q-adjusted capital stock do not at this point give more than a reason to doubt the standard finding. Even if it turns out that capital growth really was overstated in the 1970's, the policy conclusion is not obvious. If there is another wave of structural changes coming in the 1980's, it would pay to postpone major investment decisions. If the world economy has settled down, then there surely would be a case for rebuilding a seriously depleted capital stock.

REFERENCES

Kenneth J. Arrow and Mordecai Kurz, *Public Investment, the Rate of Return, and Optimal Fiscal Policy*, Baltimore 1970.

E. Berndt, "Reconciling Alternative Estimates of the Elasticity of Substitution," Feb. 1976, *58*, 59–68.

Edward F. Denison, *Accounting for Slower Economic Growth*, Washington 1979.

M. Feldstein and L. Summers, "Is the Rate of Profit Falling?," *Brookings Papers*, Washington 1977, *1*, 211–28.

B. Fraumeni and D. W. Jorgenson, "Capital Formation and U.S. Productivity Growth, 1948–1976," mimeo., Harvard University, undated.

D. W. Jorgenson, "Investment Behavior and the Production Function," *Bell J. Econ.*, Spring 1972, *3*, 220–51.

J. R. Norsworthy, M. J. Harper, and K. Kunze, "The Slowdown in Productivity Growth:

Analysis of Some Contributing Factors," *Brookings Papers*, Washington 1979, Feb. 1979, *2*, 387–421.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

G. M. von Furstenberg, "Corporate Investment: Does Market Valuation Matter in the Aggregate?," *Brookings Papers*, Washington 1977, *2*, 347–408.

Cultivation of Taste, Catastrophe Theory, and the Demand for Works of Art

By ROGER A. MCCAIN*

In the process of cultivation of taste, tastes are changed by the experience of consumption. To model this we might assume that cultivation of taste for a particular good involves a change in only one parameter of the utility function, and that the other parameters, that is, the "underlying" utility function, remains unchanged. For example, assume that there are just two goods, x and z, of which x is subject to cultivation of taste, and z is not. Then

$$U=f(Ax,z) \tag{1}$$

where x and z are consumption flows of two goods, U "utility," and A is the parameter which changes as taste is cultivated. Then A depends on the individual's history of consumption of x and z. (See Dava Sobel.) This is, of course, an application of the Becker-Lancaster approach (see Kelvin Lancaster, my 1979 article, Robert Michael and Gary Becker, Robert Pollack, and George Stigler and Becker). While we might consider cultivation of taste for goods other than works of art, we presumably must understand the economics of cultivation of taste as a first step in constructing a theory of the demand for works of art and for artistic performances.

When we allow for multiple equilibria and endogenous determination of taste, two issues arise which we need not face otherwise. First, we might assume either that consumers are shortsighted or that they are farsighted. Second, we might assume that they are calculators or that they are gropers. By farsighted, I mean a person who takes the accumulation of taste capital into account in his or her decisions. By a groper, I mean a person who seeks the optimum by a process of trial and error, which uses only local information, and which consequently may yield only a local rather than a global optimum. By a calculator, I mean a person who chooses only the global optimum. This paper assumes that consumers are farsighted gropers.

I begin the model with equation (1) above. Next I posit the determination of A over time by the flow of consumption of x:

$$\dot{A}=G(x,A) \tag{2}$$

Now, x and z are instantaneous flows of consumption, so taking z as the numeraire,

$$px(t)+z(t)\leqslant c(t) \tag{3}$$

where $c(t)$ is total consumption expenditure at instant t. Moreover,

$$\int_0^T e^{-rt}c(t)\,dt\leqslant W \tag{4}$$

where W is the discounted present value of lifetime income. The consumer is assumed to maximize the discounted present value of the flow of utility from consumption,

$$max\int_0^T e^{-rt}f(Ax,z)\,dt \tag{5}$$

subject to (2), (3), and (4), and subject also to constraints of a given initial and terminal stock of taste capital.[1]

This problem yields the following characterization for a stable "underlying de-

*Associate professor of economics, Temple University. I am indebted to Tom Sherwood and William Hendon for persistently encouraging me to explore the economics of the arts. Shortcomings of this paper are, of course, attributable to me.

[1] A more extended discussion of these matters, including mathematical details, is contained in my 1980 working paper which is available on request.

mand curve":

$$(6)\qquad \left[1-\frac{x\dfrac{\partial G}{\partial x}}{A\dfrac{\partial G}{\partial A}}\right]\frac{\dfrac{\partial f}{\partial(Ax)}}{\dfrac{\partial f}{\partial z}}=\frac{p}{A}$$

Notice that while (6) is the form which corresponds to the Stigler and Becker underlying demand curve, it is rather more complex. To obtain the simpler form,

$$(7)\qquad \frac{\dfrac{\partial f}{\partial(Ax)}}{\dfrac{\partial f}{\partial z}}=\frac{p}{A}$$

as a general form, we must assume that people are shortsighted, and that assumption seems to be implicit in the Stigler and Becker analysis.

Nevertheless, let me make some preliminary comments about the possibility of multiple equilibria in terms of (6). In the parenthetical phrase on the left-hand side of (6), a larger x and a larger A will tend to offset one another, so that the elasticity of "underlying demand curve" (6) will be approximately the same as that of (7). That the underlying demand curve be elastic is a necessary condition for multiple equilibria, since the elasticity of the underlying demand curve implies that larger A is associated with larger x (see Stigler and Becker, and my 1979 article). This may be illustrated by Figure 1, where D is the locus of stable demand (i.e., of (6)) drawn on the assumption that underlying demand curve is elastic. Curve G is the locus of $G(x, A)=0$. An equilibrium can occur only in the intersection of the two curves. Since both curves are upward sloping and presumably nonlinear, multiple intersections are clearly possible. Each intersection of D and G is a potential equilibrium, though not all need be stable.

The case illustrated in Figure 1 corresponds to that explored in my earlier paper (1979). In the context of the simpler (cobweb) dynamics of that model, the equilibria at a and c would be locally stable, and that at b unstable. That may be so in this case as well, though the local stability of a and c depends upon the exact parameters of the model. Henceforward I will assume that they are stable. Multiple equilibria are not the whole of catastrophe theory, however, (see C. Zeeman, Hal Varian, and my 1979 article). We might have multiple equilibria and not have discontinuous change. We must now ask whether continuous variation in the parameters of the optimization problem, p and W, can cause the disappearance of some locally stable equilibria and a consequent discontinuous movement of the consumer to another of the locally stable equilibria. (In this context "discontinuous" movement simply means movement out of equilibrium, by contrast with the movement of the equilibrium itself consequent on the continuous variation of the parameters.) In Figure 1, for example, a decrease in p would displace D rightward, and a sufficiently large displacement clearly could eliminate the equilibrium at point a, leaving only that at point c. Thus a consumer formerly in equilibrium at point a would move discontinuously (out of equilibrium) to a new equilibrium at point c.

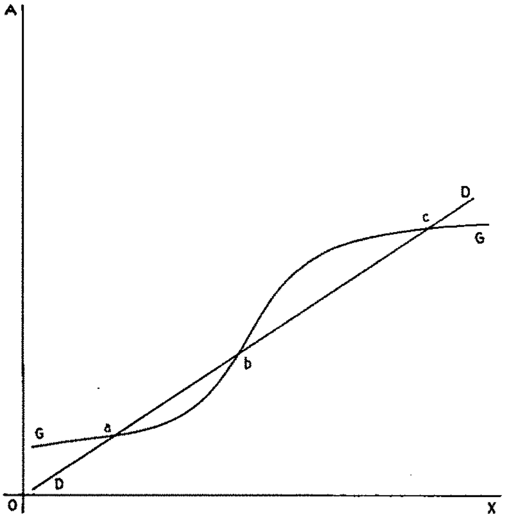


FIGURE 1

This authorizes us to consider, at least as a possibility, a supply and demand model such as that shown in Figure 2. In Figure 2,

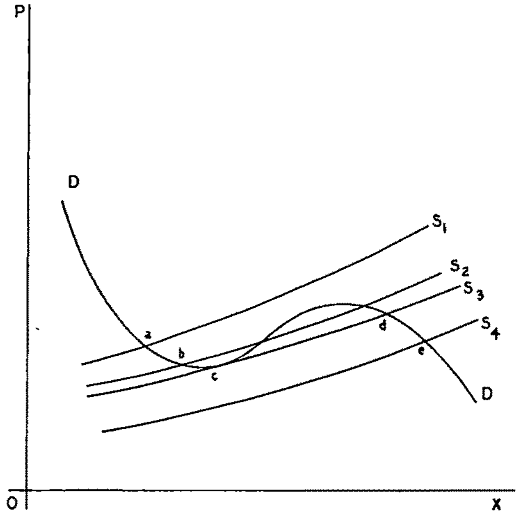


FIGURE 2

D is the long-run demand curve and S_1, S_2, S_3, and S_4 are successive supply curves. The consumer begins at point a, follows the gradual rightward shifting of the equilibrium to point c, shifts discontinuously to point d, and then again follows the gradual shifting of the demand curve downward, in equilibrium, through e and beyond. This has apparently occurred in the market for table wine in the United States. Remarkably, even if the underlying relations are linear, D in Figures 1 and 2 must be at least of the third order.

A second empirical implication has to do with cross-sectional distributions of consumption of goods subject to cultivation of taste. Under stable conditions of supply and demand, one might expect that the consumption of the good would be distributed across the population in a multimodal fashion. In Figure 1, which assumes a given price and income (or wealth, W), if individuals are quite similar in their underlying demand or utility functions and in their capacities to accumulate taste capital, then we would expect to find them clustered around a and c, and few if any at b. In saying that I presuppose that a and c are locally stable.

Moreover, we may expect demand catastrophes as income or wealth increases, just as we might expect them with changing prices. If x is a normal good, an increase in income or wealth would tend to shift D in Figure 1 rightward. Locally, this would displace a and c rightward, and b leftward; it could also lead to the disappearance of the lower A equilibrium at a sufficiently high income, so that from that income level onward, only the higher level equilibrium would exist.

These possibilities are not only observable in principle, but observable cases are likely to be far less rare than demand catastrophes. The data necessary to look for them may well be available. They have not been sought, and empirical work is beyond the scope of this paper, but this seems a promising area for the experimental application of catastrophe theory to the demand for works of art and for artistic performances.

REFERENCES

Kelvin Lancaster, *Consumer Demand*, New York: Columbia University Press, 1971.

R. A. McCain, "Reflections on the Cultivation of Taste," *J. Cultural Econ.*, June 1979, *3*, 30–52.

______, "Cultivation of Taste, Catastrophe Theory, and the Demand for Works of Art," work. paper, dept. econ., Temple Univ. 1980.

R. T. Michael and G. S. Becker, "On the New Theory of Consumer Behavior," *Swedish J. Econ.*, Dec. 1973, *75*, 378–96.

R. Pollack, "Habit Formation and Dynamic Demand Functions," *J. Polit. Econ.*, July/Aug. 1970, *78*, 745–63.

D. Sobel, "Acquired Tastes Found to Have Survival Function," *New York Times*, Mar. 4, 1980, p. C1.

G. Stigler and G. S. Becker, "De Gustibus non est Disputandum," *Amer. Econ. Rev.*, March 1977, *67*, 76–90.

H. Varian, "Catastrophe Theory and the Business Cycle," *Econ. Inquiry*, Jan. 1979, *17*, 14–28.

C. Zeeman, *Catastrophe Theory*, New York: Addison-Wesley, 1977.

Economic Theory and the Positive Economics of Arts Financing

By BRUCE A. SEAMAN*

The essential question for a positive theory of arts financing is: what determines the quantities and relative shares of the different sources of arts support? These sources include the earned admission revenues from purely private financing, and the "unearned" primarily lump sum money grants given by individual private contributors, corporations, foundations, and governments at all levels.

Unfortunately, our analytical understanding of arts financing is still plagued by anomalies. For example, empirical studies of nonfederal government support have generally found income variations to be only weakly related to arts grants (see Dick Netzer, p. 52), or have found evidence that communities with a larger proportion of very high-income people compared to moderately high-income groups actually provide less local government arts support (see my 1979 article).

This paper examines a particular government arts subsidy study and asserts that 1) direct adaptation of successful government expenditure models to the arts case requires special precautions to avoid misinterpreting the results, especially in the estimation of the "publicness" of the arts; and 2) the interaction between public and private unearned arts income may partly explain the behavior of income in previous subsidy studies.

I

Using Australian state expenditure data, Glenn Withers found median income, population density, tax share, and federal arts outlays to be important determinants of state support, but found little impact of an "upper-income" variable (the proportion of taxpayers with above one and one-half the median income) expected to be particularly important in generating public support for the arts. It was therefore argued that the importance of median voter demands for political decision making is confirmed, and that contrary to popular assumption, it is not the wealthy who dictate arts policy (p. 59). Furthermore, a "crowding parameter" considerably greater than one was estimated, leading to the conclusion that the arts have few public good characteristics, so that public arts expenditures probably reflect "private gains accruing to the median voter group" (p. 58). However, a more careful examination of this model suggests other possible interpretations, and highlights some difficulties in designing adequate arts funding models.

The model utilized by Withers defines the quantity of the publicly financed good as Z, and the relationship between Z and the units actually consumed by any representative individual in the community as $Z_i = Z/N_c^\gamma$, where Z_i is units of Z consumed by person i, N_c is the size of the population actually sharing in the consumption of Z (which may be less than taxpaying population, N), and γ is the crowding parameter equal to zero if the good is purely public in the sense of being able to be equally consumed by all N, and equalling one if the good is fully divisible in consumption (see Theodore Bergstrom and Robert Goodman). If we further define t_i as i's tax share, often designated as $Y_i/\Sigma_{i=1}^N Y_i$ (i's taxable income, or assets relative to that of the community), the personal expenditure per actual unit consumed is $t_i qZ/(Z/N_c^\gamma)$, or simply $t_i qN_c^\gamma$, with q equal to the per unit cost of good Z. This expression represents i's per unit price of Z_i financed publicly. Alternatively, if the good can also be financed privately, the consumer could presumably buy Z_i at unit cost q.

*Assistant professor of economics, Georgia State University.

The relationship between Z and Z_i suggests that any constituent demanding Z_i must express a demand for the public financing of $Z_i N_c^\gamma$. Therefore, i's demand function for units of Z (assuming $N_c = N$) is $cq^\delta t_i^\delta Y_i^\varepsilon N^{\gamma(1+\delta)}$, with δ the price elasticity, ε the income elasticity, and c a constant. The coefficient of population estimated in a *log*-linear regression on public expenditures thus allows for the computation of the product's publicness, γ.

What product is being financed when the government makes expenditures on the arts? Commonly used definitions of arts output include number of performances or gallery showings, number of actual tickets sold, or even number of arts "experiences" consumed.

The argument to follow is that if number of performances or actual tickets sold is designated for output, the crowding test fails. Furthermore, there is no logical reason for using these output measures, since in the arts case the government does not really provide either of these things directly. The government and the other sources of unearned income provide lump sum money grants to be used by the recipient organizations as they see fit. Sometimes these grants lead to wage rate and employment changes. Other times they increase the quality and variety of arts events without appreciably affecting the quantity. Most frequently it is thought that they lead to reduced ticket prices and increased attendance at existing performances, a result depending on the recipient firm's costs, seating capacity, and the price elasticity of demand for tickets.

It is therefore less misleading to view the product most directly provided by the government as simply money grants. After all, individuals in the community are not generally faced with a decision of whether to buy arts output from private firms or from the government. They almost always buy it privately from nonprofit organizations, with the quality and price of the product dependent upon their willingness to "collectively" finance (in the sense of an implicit community cost sharing arrangement) supplemental funding from either public or private sources. The private substitutes for public provision are clearly the private money grants potentially available from individuals, business firms and foundations.

For example, assume Z is called number of arts performances or events. I have argued elsewhere (1980a) that applying the narrow interpretation of Bergstrom-Goodman to the notion of representative individuals sharing in the consumption of Z (where no discretionary spending is involved, just technical limitations on the collective consumption of the product), a plausible relationship between Z_i and Z would be $Z_i = (K/N_c)Z$, where K is the maximum capacity of arts facilities. One interpretation would be that the probability of an individual consuming any one performance, or the individual consumption of performances relative to total performances is a positive function of both the viewing capacity relative to the arts audience, K/N_c, and the number of performances actually offered in any time period. The only way that we can get the fully private good case is for capacity K to equal one, and the only way for the pure public case is for $K = N_c$. Since these extreme cases are almost inconceivable, an attempt to estimate the degree of publicness of arts performances with a crowding parameter will be misleading. Furthermore, since an individual's price per unit of Z collectively financed would be $t_i qZ/Z_i = t_i qZ/(K/N_c)Z$, or finally $t_i q(N_c/K)$, we don't get the tax price per unit, $t_i q N_c^\gamma$, necessary to apply the Bergstrom-Goodman crowding test.

However, this dilemma is perfectly understandable, since the typical Bergstrom-Goodman case of trying to estimate the degree to which police services, parks and recreation, or garbage collecting can be jointly consumed by community residents is quite different from the arts case. Arts services are not only fully excludable from nonpayers (except for consumption externalities, which is a separate issue not considered by Bergstrom-Goodman), but rationing by price is actually used. The other services are essentially examples of full public financing. The number of arts performances consumed by any individual, Z_i, is fully dependent upon the discretionary spending of that individual for admission tickets, which is a function of admission prices, substitute good prices, income, value of time, and tastes.

Therefore, a better measure of individual consumption, Z_i, would be actual tickets sold. But again, strict application of the crowding framework is impossible. If Z is the number of tickets sold, then Z must equal $\Sigma_{i=1}^{N_c} Z_i$, or $Z = N_c\bar{Z}_i$, where $\bar{Z}_i$ is the average number of tickets sold to arts consumers. Again following through with the analysis, $\bar{Z}_i = Z/N_c^{\gamma}$. But of course $\bar{Z}$ by definition is the average number of tickets sold to consumers times the number of consumers, so that γ must by definition equal one. Obviously this is the case, since full crowding exists in the consumption of any seat at an arts performance. And since the public price per unit for the average consumer is $t_i qZ/(Z/N_c)$, or simply $t_i qN_c$, we have the special case noted by Bergstrom and Goodman (p. 282) in which we should consider demands for *per capita* expenditures as a function of income and the "tax price multiplied by the population" (here the arts consumer population). This is different from the crowding test case of estimating demand for total expenditures as a function of t_i, income and a separate population variable, with the estimated population coefficient allowing for the derivation of the degree of publicness of the product.

On the other hand, if Z is defined as money grants, the degree to which individuals in the community privately consume these grants, in the form of "effective grant dollars," would again be expressed by $Z_i = Z/N_c^{\gamma}$. And to whatever extent an effective grant dollar depends on the size of the community, (i.e., $\gamma \neq 0$) we get a kind of crowding in the consumption of grants. Such crowding is apt to be present since, if it were not, it would be true that a dollar given in a large community is equally effective as one given in a small community. But this is unlikely when a major expected benefit of the grants is the lower admission fees that result.

As noted, this will depend on costs and arts capacity (or size of the target audience). For example, assume two local cultural institutions in different communities have the same "average admission cost," equal to the total operating cost (for all performances given) divided by the maximum audience that could be serviced, given seating capacity. If demand is such that a capacity audience can be obtained by charging prices equal to the average admission cost, and these nonprofit organizations are satisfied charging that price so as to just break even, the larger organization will not reduce prices as much as the smaller in response to an equal-sized grant. This is just arithmetic. If for the larger, \$100,000/10,000 = price per ticket of \$10, and for the smaller, \$10,000/1,000 = \$10 price, the receipt by each of a \$1,000 grant will reduce the former's price by 10¢ (\$99,000/10,000), and the latter's price by \$1.00 (\$9,000/1,000). Since this implies that the effectiveness of \$1 of grant money is inversely proportional to the "scale" of the arts sector, taxpayers interested in buying \$1 worth of private benefits from publicly financing the arts must stand willing to finance more grants as the size of the audience (or capacity) increases. This again means that the per effective grant dollar price to an individual taxpayer is $t_i qN_c^{\gamma}$, or just $t_i N_c^{\gamma}$.

If the value of γ were one, it would justify considering the demand for grants "per arts patron," or per capita for the community if $N = N_c$. Actually, γ could be less than one if recipient organizations are able to make particularly good use of grants in reducing prices (or providing anything else of value to taxpayers) due perhaps to economies of scale that would allow the larger organization above to reduce price by more than 10¢. On the other hand, there would be a logical interpretation to a crowding parameter greater than one (i.e., in order to obtain \$1 of effective grants, taxpayers must finance grants more than proportionally larger than arts consumer population, N_c).

Assuming taxpayers only valued increased arts attendance due to reduced admission prices, this γ would be greater than one if, say, musicians unions were able to induce arts management to increase employment or wage rates as grants increase.

Or it could result if price elasticity of attendance demand is low (as indicated by empirical studies), making it difficult to obtain additional attendance even if prices fall substantially.

In this context, Wither's estimate of the crowding parameter as greater than one could mean that arts *grants* are crowded (and therefore not pure public goods) in the sense that the individual private benefits perceived by the taxpayer demanders vary inversely with population size so that more grants are required in larger communities than in smaller communities to have the same effective result. It says nothing about the publicness of the arts themselves.

But this does not necessarily imply that public outlays for the arts have little to do with attempts to "capture public good characteristics not obtainable through market demand" (Withers, p. 58). It is not uncommon to view goods that are basically private in the sense of crowding in consumption as having sufficient externalities so as to treat them as public goods. This is done consistently in the study of the demand for public education spending, where the product is defined as "units of education per pupil," or "dollars per unit per pupil."

Similarly, arts grants, whether financed publicly or privately (through the private philanthropy "market") yield externalities that cannot be fully withheld from nonpayers. The price reductions and quality and quantity changes that result from such grants are "available" to anyone in the community. It is true that in this case these external benefits are more likely to accrue to a definite subset of the population (arts lovers). But for that particular community, the problem of optimally providing arts grants (primarily free riding) is similar to that of broader communities trying to optimally provide more widely consumed public goods.

Since there is a systematic relationship between population size and the difficulties of free riding, the population variable used by Withers would still be important, but with a different interpretation. Its strongly positive effect could possibly mean that the arts "interest group" looks toward government support when the per capita private support drops. Since the coefficient on population is greater than one, meaning elasticity greater than one, per capita expenditures increase with population. Further confirmation of this interpretation would require results also on per capita contributions. In my earlier 1980b study, the *log* of population size of the *SMSA* was found to be strongly positive in determining the *log* of per capita government grants, but either negative or weakly positively on per capita *SMSA* contributions (but not statistically significant). This suggests some support for this interpretation.

II

Furthermore, there is a difference between the hypothetical case of community-wide public goods and this case of art lovers public goods: the taxpayer community N is larger than the arts consumer population N_c. This is similar to the case of public secondary education, where the direct consumers are fewer in number than the total taxpayer population. In both cases, the consumers will face a lower price for the product via collective financing than with pure private financing, since part of the costs of production are being paid by nonconsumers.

But in contrast to the secondary education case, where the form of collective financing is to have the government collect taxes and directly run the schools, collective financing of arts grants can be done either publicly *or* privately. Thus, while arts consumers will prefer some form of collective financing to pure private financing, their choice among types of collective financing, and the impact of their particular demands on the overall community political decision to publicly finance arts grants, is problematical.

Private substitute prices relative to individuals' tax prices can be used to determine how the quantity of public expenditures demanded varies with personal income. Since richer taxpayers have higher real incomes, but face higher tax prices of publicly financed goods relative to the prices of private substitutes, than do lower income taxpayers, quantity demanded will increase with income only if the income elasticity of demand exceeds the "elasticity of substitution between collective goods and privately consumed goods" (see Lawrence W. Kenny, p. 117).

Thus the relationship between personal income and public spending demanded could take the form of an inverted U, with middle-income groups demanding more than either very low- or very high-income groups. This could happen if the income elasticity of demand is outweighed at very high incomes by the elasticity of substitution between publicly and privately supplied goods. The probability of this happening would be enhanced if, as Burton Weisbrod has argued, the income elasticity of public demand is itself lower at higher incomes since private substitutes generally yield more "individual control" by the buyer than do publicly provided services. If such control is a normal good, the demand for the private good will rise with income relative to the demand for the public substitute.

In the arts, if we focus only on personal tax deductible contributions as the private sector alternative to public grants, not only will the public sector tax price for potential contributors tend to rise with income, but the price of contributing privately will fall with increases in income and the marginal tax rate. Thus, the chances of the income elasticity of public grant demand being outweighed by the elasticity of substitution between public and private grants may be higher than in cases where private prices are independent of income.

Furthermore, a particular version of Weisbrod's varying income elasticity will exhibit itself in the arts case because as income increases, potential contributors may prefer to give privately rather than approve of being taxed further to support public grants. If the household production approach to consumer behavior is applied to the production of cultural experiences by consumers, one can view privately given grants as reducing the shadow price of a unit of arts experiences, since being a contributor often includes certain privileges that may make arts attendance more "productive" (i.e., better seating, invitations to special showings and backstage introductions, etc.). Or it could be viewed as reducing the shadow price due to the reductions in the amount of time needed to produce a unit of arts experiences when contributors are allowed to skip the lines at the King Tut exhibit, or due to other time saving perquisites conferred upon contributors. In this sense privately contributed grants confer privately appropriately benefits as well as public good externalities. Thus, for given relative prices of giving privately and publicly, potential donors will want to "buy" more private arts grants as income increases.

Alternatively, we could view this in light of the crowded goods discussion. Privately given grants are less crowded to the individual donor than are publicly "given" grants. That is, for any given grants expenditure, more effective grant units can be actually consumed (since the effectiveness of the private grants is not limited to their effects on, say, ticket prices), and therefore the price per unit of effective grant, Z_i, is reduced.

These considerations suggest an alternative interpretation to Withers' upper-income variable. Its poor performance relative to median income in predicting cross-sectional variations in Australian state arts expenditures might be due to the inverted U relationship between income and the quantity demanded of publicly financed arts grants. A similar explanation has been offered for the success in subdividing an upper-income variable into a "moderately" high-income proportion of the population, and a "very" high-income proportion of the population when trying to explain cross-sectional government museum grants (see my 1979 article). Netzer's problem with finding no effect of income in *U.S.* state arts expenditures may also be partly due to the complexity of the relationship between demand for publicly financed grants and income, when private grant substitutes are taken into account.

Improving the quality of arts data remains perhaps the primary problem in arts research. This paper has attempted to demonstrate that careful model construction is also important in developing an understanding of arts financing, and interpreting the empirical results that have already been obtained.

REFERENCES

T. C. Bergstrom and R. P. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63,* 280–96.

L. W. Kenny, "The Collective Allocation of Commodities in a Democratic Society: A Generalization," *Public Choice,* 1978, *33,* 117–20.

Dick Netzer, *The Subsidized Muse: Public Support for the Arts in the United States,* Cambridge 1978.

B. A. Seaman, "Local Subsidization of Culture: A Public Choice Model Based on Household Utility Maximization," *J. Behavioral Econ.*, Summer 1979, *8,* 93–131.

______, (1980a) "Private Demand for Public Subsidies: A Comment," *J. Cultural Econ.*, June 1980, *8,* 47–54.

______, (1980b) "Economic Models and Support for the Arts," in W. S. Hendon et al., eds., *Economic Policy for the Arts,* 80–95, Cambridge, Mass. 1980.

G. A. Withers, "Private Demand for Public Subsidies: An Econometric Study of Cultural Support in Australia," *J. Cultural Econ.*, June 1979, *3,* 53–61.

B. A. Weisbrod, "Toward a Theory of the Voluntary Non-profit Sector in a Three Sector Economy," in Edmond S. Phelps, ed., *Altruism, Morality and Economic Theory,* New York 1975, 171–195.

Supply Decisions of Professional Artists

By LESLIE P. SINGER*

At a sale at Sotheby's on October 18, 1973, a "Double White Map" by Jasper Johns, a living American artist, was sold for $240,000. Was it rational for Johns to let Robert Scull acquire the work in 1958 for $10,200? A more recent transaction involved the sale for a reputed three million of Johns' "Three American Flags," originally acquired for $500. We have records of similar sales of Pollocks, de Kooning, Warhols, etc. Surely Johns could have made more flags in his inimitable style. Each, if not an exact duplicate, at least similar in composition and spirit, yet no great value would be attached to such products.

Artists' supply decisions differ from those of other self-employed professionals, say, dentists or physicians, inasmuch as the works are storable by both the producer and consumer. Speculative capital gains (or losses) are realizable by both. The use of home time is intertemporally transferable and thus subject to stochastic optimization.

If the principal source of consumer utility is the objective characteristic of craft (composition, etc.) which was the case with Renaissance artists, differences in market value between an artist's early and late works tend to be due to constraints on the information set. On the other hand, if the source of utility (for a connoisseur) is invention, significant disparities exist in market values between early seminal works and late works.

A dramatic development has taken place in art appreciation, dating from about the beginning of the current century. The implications were both aesthetic—in a broad cognitive sense—and economic. What emerged was an "action theory" of value. The fundamental tenet is that art historical significance resides in a nonreplicable act of creation. Possibly the first application of an action theory occurred when early Fauve works by Derain, Comoin, Friesz, and others of the short-lived movement suddenly acquired art-historical significance far beyond their aesthetic merits, while later works, executed with superb craftsmanship were relegated to lower price ranges. Similar distinctions were made of surrealist works by de Chirico, late expressionist works by Schmidt-Rottluff and Nolde, or late action paintings by de Kooning and current pop works by Warhol.

I specify art as a bundle of Lancastrian characteristics of "decorativeness" and "intellectual appeal" for which there exists a consumer technology (see my earlier article, pp. 21–22). Decorativeness comprises objective characteristics: color, composition, draftsmanship, pigmentation, brushwork. Some invention and or originality can be subsumed under decorativeness. Examples are the luscious coloring of Rubens or the dramatic spatial organization of Raphael. Intellectual appeal, while often conjoint with characteristics of decorativeness, establishes art historical significance, and often precedes public discovery of "decorativeness." This is particularly true of twentieth-century art where a discerning aesthetics, that is, a consumer technology capable of generating satisfaction from stimuli such as the loose pigmentation of the Impressionists, the glaring color contrasts of the Fauves, or the ramshackle assemblage of the environmentalists, was not in place when the works were first marketed.

Let us assume the artist is a natural monopolist seeking to maximize not profit but a utility function in income Y, leisure L, and own art x, subject to a time constraint, certain market constraints, and a simple technical production relationship containing the artists' own time H, and possibly other inputs such as assistants, equipment, etc.

Contemporary art's preoccupation with authenticity places a historically novel emphasis on time-specific personal execution. Rubens, Raphael, and most other

*Indiana University.

classical artists employed specialized labor; some workshop artists painted draperies, others fruit, yet others pricked cartoons. The great Claude Lorraine was not loath to letting others paint figures in his landscapes. Vasarely and Warhol are among a few contemporary artists who tried the age-honored method of production by skilled labor and have done so with varying success. The art establishment prefers works by the artist's hand. Thus we can assume without loss of generality that factor costs are zero.

Let the artist's utility function contain leisure L, goods q, and own art x. Assume no taxes and no savings; that is, the market value of sales (barter) equals consumption. Given unearned income (studio sales), the value of own time in market activities, $w(h)$, and a price vector for $n-1$ commodities p_i, we can define labor supply and commodity demand functions in the usual manner.

Suppose the artist rations x out of the market. Let p_n be the shadow price of x, that is, the wage that would have been earned had x been sold. The artist's equilibrium supply of his own time, under rationing, can be determined as in Roger Latham (pp. 307–10).

In a more conventional vein, consider Figure 1. On the horizontal axis we measure consumption of L, from west to east and work in the opposite direction. On the vertical axis, we measure income realizable through the artist's sale of his own works, which can be converted to an hourly wage w, as a function of the proportion of time spent on market activities, h. The budget curve Y reflects the fact that most artists can sell their works only at rapidly decreasing prices. I shall assume that the artist is initially in equilibrium when he engages in h_2 market activity and does not retain any x. If the artist retains some x, his budget curve will be Y'. The distance between Y and Y' is the opportunity cost $p_x = dY$, corresponding to each level of market activity; namely, the marginal loss of income which the artist sustains by consuming an additional unit of x, instead of raising h and thus consuming an additional unit of q. Clearly, p_x will decline as market activity h expands; the rate depending on the elasticity of Y: $p_x \to 0$, as $e \to \infty$.

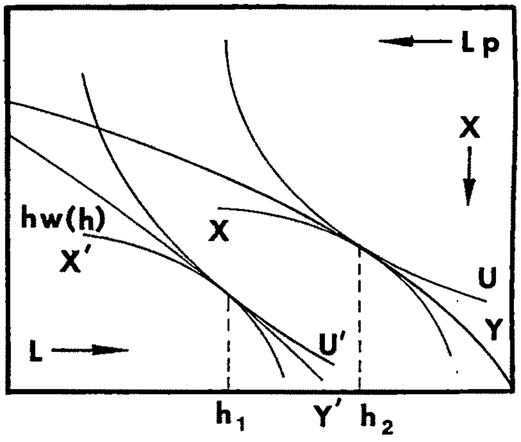


FIGURE 1

Consider now the MRS of x for q or, equivalently, of x for work Lp, as represented by X and X'. Figure 1 is normalized in such manner that at h_2 where X is tangent to Y, no own art is consumed. At h_1, $(h_1 - h_2)$ work is converted into income and h_2 into consumption of art, x, at an opportunity cost which equals the difference between the two points of tangency, smaller than $Y - Y'$ at h_1. The artist is in equilibrium by moving lower on the $U(q)$ curves from U to U' but correspondingly moving up from X to X'. I assume that both q and x are normal goods, thus more x as well as more q will be demanded at higher incomes. The cost of storage is ignored (notwithstanding that Picasso had to move from chateau to chateau in order to accomodate his voracious appetite for x).

It is reasonable to assume that innovative artists confronted with an underdeveloped consumer technology, would be governed by more elastic constraints than imitative artists. Consequently, the opportunity cost of home consumption of art would be much lower for innovative artists. Other things equal, the innovative artist will hold more own art for identical h than an imitative artist who tends to hold no x. Furthermore, the innovative artist will consume less leisure or work more than the imitative artist, even though the imitative artist's market activities tend to exceed those of the innovative artist.

The artist in secondary markets of vintage art realizes that he is supplying a dated product, say, q_{ij}, namely art of vintage i sold in period $j, i \leqslant j$. Each period the artist enters with a new vintage and the market can choose at any period, j, among q^*, a vector of vintages, $i<j$. In addition, collectors, galleries and auction houses can supply in period j any vintage $i<j$, purchased in period k, $i \leqslant k \leqslant j$. The artist can withhold any vintage i from reaching the market in any $j>i$.

It is tempting to compare the revenue-maximizing process engaged by the artist, who is a primary seller, confronting secondary market competition, with a monopolist such as Alcoa, confronting a secondary market of recycling firms with given recovery costs and primary and secondary product prices as in Peter Swan (1972, 1980). In the case of fad art, only a fraction of the original value may be recovered in successive secondary sales and price maintenance may involve extensive promotional costs (similar to recovery cost in recycling). However, unlike the Alcoa case, most nonimitative art tends to first rapidly gain in secondary markets, then stabilize and subsequently decline, as new entries overtake fading "schools."

Consider Figure 2, where q_{11} is current art sold in the present while q_{12} is current art scheduled for sale in an indefinite future period. We may assume that the artist can borrow unlimited amounts against future sales (or holds another job). Let future sales be fully discounted. Assume zero costs and artist's inverse demand functions, $p_1(q_{11})$, $p_2(q_{12}, q_{22})$ etc. with property that the partials with respect to q_{ij}, for $i<j$ may be negative or positive and the cross-price derivatives $\geqslant 0$. Strategic sales by the artist of certain vintages may complement sales of current vintage art, often spectacularly raising the prices of the artists' works of all vintages (and of diverse quality.)

Let us consider two possibilities. First consider the case where there is no secondary market. The artist brings about a division of output into present and future sales, as in Heinrich von Stackleberg. The relevant marginal revenues MR_1 and MR_2 $\geqslant 0$, are equated as in Figure 2 and the artist simultaneously determines his total output of vintage i art as well as each period's sales. For outputs below q there are no current sales. The artist borrows then repays the loan in the next period. He sells q_{12} as well as the new vintage q_{22}, thus maximizing total net revenue. Next consider a secondary market by collectors. Again strategic sales can raise prices, whereas dumping of the artist's works by collectors, who change trends by espousing new fashions, has the opposite effect. The separation of markets is now given as in Richard Cebula. Moreover, q_{12} is no longer under the artist's sole control and is a function of prevailing prices and sales by both artist and collectors. Record prices achieved by some vintages will not only shift forward the artist's own demand curve for other vintages, but will also attract into the market—the auction houses—collectors who had previously made strategic purchases from the artist. This affects market demand somewhat along the lines of Swan (1980) or Darius Gaskins.

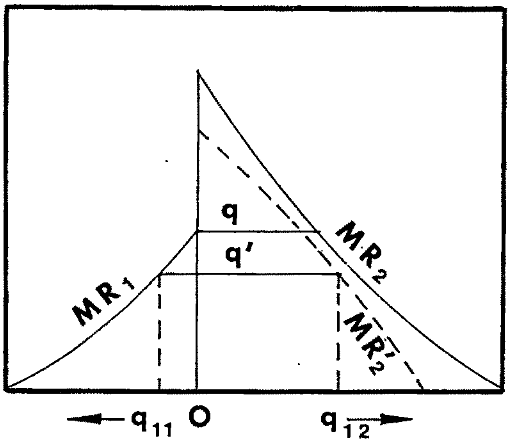


FIGURE 2

In a market where the artist's primary sales v equal total sales q, less secondary sales $s, s=f(v_{-1})$, one can use Swan's (1980, p. 83) formulation of long-run equilibrium for each vintage; namely, $p(q^*_{ij})(1-e'/e)$ which must be equal for all j, i constant. Here e is conventional price elasticity, e' is $(dq/dv)/(v/q)$. There is an implicit assumption that the artist, having decided how much of a given vintage to produce and sell (as in Figure 1) will not subsequently change his mind. This is equivalent to the constant supply of scrap assumption in Swan (p. 81). Clearly, if there are no secondary sales, (i.e., no recovery) $dq=dv$, $e'=1$. $MR_1=MR_2$, as in Figure 2. If $e'>1$, the usual case for nonimitative art held speculatively, MR' replaces MR. The interesting conclusion follows that now optimization involves shifting some q_{12} to q_{11}, as shown in Figure 2. The reverse is true if $e'<1$, when $MR'>MR$. The initial impact of a record price, such as for John's "Three American Flags," induces secondary sales. Subsequently, the rate of appreciation of all vintages diminishes and long-run equilibrium output as well as the optimal division between current and future sales is established, as in Swan (p. 83), where $-dv=ds$ $e'=0$ and $MR=AR$. The artist or his dealer, makes a complete adjustment to secondary sales. On the other hand, artists such as Chagall or Picasso, can sell all they produce at a given price; clearly $e\to\infty$ as $e'\to 0$.

Increasingly more complex models can be built to explain multistage processes. The analogy with recycling no longer holds unless couched in dynamic programming terms. An important difference, possibly unique to art, is that the artist can choose to enter competitive or atomistic markets and, in fact, the technology exists for him to operate in both markets simultaneously.

I. The Quality-Quantity Spectrum

Suppose the artist, given his imagination and skills, can distribute characteristics of "decorativeness" and "intellectual" appeal" on a quantity-quality continuum. For the same input of leisure (or work) he can choose to trade quality for quantity by more heavily weighting characteristics towards "decorativeness." This choice can be made at any time and in any sequence.

Product-mix decisions involve choice of markets. As one shifts from quantity to quality, one moves from purely competitive markets with freedom of entry and near-perfect knowledge to imperfect markets with uncertain prices, significant entry barriers, and imperfect knowledge. The artist's decision depends not only on his human capital endowments but also on his attitudes towards risk. He has to balance high potential rewards, attainable with relatively low probability, against low potential rewards, attainable with high probability. I index art by a, $0<a<1$, defined as the proportion of embodied quality characteristics: $a(1)$ stands for highest quality, $a(0)$ indicates absence of quality; namely characteristics of pure decorativeness. Suppose the artist chooses to maximize the present value of his utility stock. Let this involve the choice of some control function $g(a)$ such that $g(1)$ is associated with the highest growth rate of life cycle earning capacity while $g(0)$ is associated with minimal growth. Let market activity $h\to 0$ as $a\to 1$.

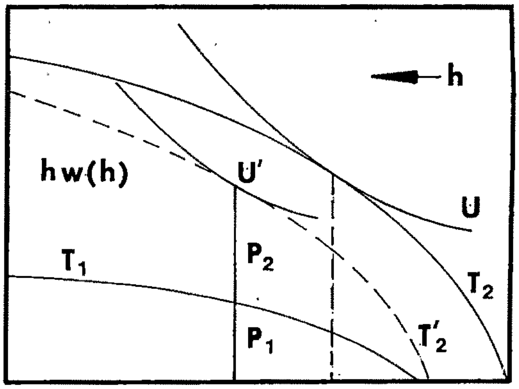


FIGURE 3

In Figure 3, T_1 and T_2 represent all possible revenue-maximizing distributions of sales between current and future periods (or series of subperiods) for given $g(a)$, e, and e'; U and U' are as in Figure 1. Thus P_1+P_2 is present value of market activities h; P_1 represents current earnings. The closer $g(a)$ is to $g(1)$ the greater will be the divergence between the artist's potential (or future) sales P_2 and his current sales, P_1. We can think of $g(1)$ as maximum investment in the artist's reputation, accomplished in the traditional manner by sacrificing current consumption for future consumption. Clearly, however, the artist must make some sales, even if future sales were more profitable, because otherwise the information set would be empty. Sacrifice sales to tastemakers (or sales below the opportunity cost of future sales) is often the price of filling the information set. If Johns had not sold any of his early work, no records would have been established, T_2 would have intersected the southwest corner of Figure 3. The artist may choose eternal bliss for posterity and martyrdom in the present, then T_1 approaches the northwest corner; and X' veers towards tangency at the southwest corner. This is the case of the "Happy Starving Artist." Typically, dealers price potential $g(1)$ artists much above market prices of comparable new works with a view to minimizing current sales P_1 in favor of future sales P_2.

The purely decorative artist derives no utility from own art and produces only art he can sell, his Y curve is a straight line. Utility stock is maximized if current consumption is maximized. The innovative artist, having initially chosen $g(1)$, at some point will have to shift toward $g(0)$, so as to maximize his utility stock. One can, with minor modifications, resort to the capital-theoretic variational argument of Alan Blinder and Yoram Weiss (hereafter B-W).

Most B-W conclusions apply to artists. The life cycle production function is concave as sales of appreciating early vintages diminish current work effort (to zero for some artists such as Marcel Duchamp). For most artists $g(a) \rightarrow g(0)$ at least once. Analogously, "cycling" can occur only when the gross rate of return to the stock of own art held exceeds the discount rate or the impatience rate, as in B-W (p. 460). That such should be the case is intuitively plausible and empirically observable. Numerous artists (Picasso, Chagall, duBuffet, de Kooning, Poons) have changed styles, concentrated on graphics and multiples, corresponding to a downward adjustment in $g(a)$. Others faded away as imitators of their earlier styles.

In Figures 3 and 4, movement towards $g(1)$ shifts T_2 north and T_1 south, thus increasing the gap between present and future sales. Movement towards $g(0)$ has the opposite effect. $g(1)$ also tends to shift Y' away from Y as more x is consumed, or held for posterity (the transversality set). As $g(a)$ approaches $g(0)$, Y' approaches Y; and T_1 approaches T_2 as in Figure 4. As in all variational problems it is difficult to optimize switch time from $g(1)$ towards $g(0)$ if terminal time (death) is uncertain.

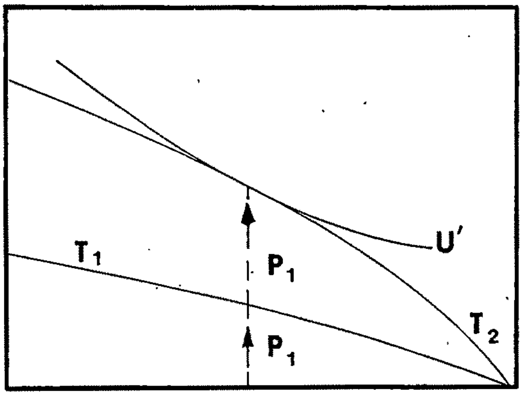


FIGURE 4

When there is general accessibility to the artist's work, that is, when the consumer technology exists, artists simply repeat their successful pieces with minor variations and no longer produce original work. The artist is over the mountain so to speak.

II. Conclusions

I have attempted to show how diverse aspects of economic inquiry can be molded into a coherent theory of art markets. One can accomplish this end by viewing art as a bundle of Lancastrian characteristics and by specifying an "action" theory of art appreciation.

I note that $g(a) > 0$ particularly $g(1)$ must be translated to the market by syndicates of secondary sellers. It would be hard to imagine the emergence of a solitary de Kooning or Pollock without the existence of a "New York" School. The market must resolve a sequence of noncooperative games where utility-maximizing artists choose strategies which strive towards an early resolution of the information set, while buyers tend to choose paths which delay the resolution, sometimes beyond terminal time.

From the artist's vantage point any movement towards $g(0)$ will increase the probability of early resolution of the information set and a competitive solution, contrariwise movement towards $g(1)$ raises transaction costs and has the opposite effect. Consequently, $g(0)$ has stochastic dominance over $g(1)$, as defined in Benjamin Eden (p. 142). The artist chooses $g(1)$ only if the expected present value of sales exceeds that of $g(0)$. Efforts by syndicates to extract monopoly rents by containing the unorganized continuum are partly responsible for the dealer-supported plethora of art movements: op, pop, top, funk, environmental, new-realism, soft and hard edge, phantastic, conceptual, etc. Thin dealer markets affect the higher moments of the distribution of talent-outcome functions by right-skewing otherwise normal populations.

Syndicates may become disadvantageous for some artists under conditions similar to Masahiro Okuno, Andrew Postlewaite, and John Roberts. Such artists may extend product lines, de-emphasizing unique works, which must be sold through syndicates to a small number of collectors, instead concentrating on graphics and multiples for which broad competitive markets exist. Artists as primary sellers face fluid coalitions. When stocks in the hands of secondary sellers increase, artists gain from the clearing of information channels. Secondary traders maximize returns on their "brands" by maintaining entry barriers, excluding nonpedigreed competitors. When primary markets contract, as secondary holdings rise, new entrants are sought out by syndicates. Secondary traders have no incentive to sustain entry barriers if the yield on existing portfolios is exceeded by returns realizable on entrants with smaller secondary holdings. Thus, each syndicate (school) dissolves when it can no longer secure a winning hand.

REFERENCES

A. S. Blinder and Y. Weiss, "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ.*, June 1976, *84*, 449–72.

R. J. Cebula, "Marked Interdependence and Third-Degree Price Discrimination: Comment," *Quart. Rev. Econ. Bus.*, Summer 1980, *20*, 106–09.

B. Eden, "Stochastic Dominance in Human Capital," *J. Polit. Econ.*, Feb. 1980, *88*, 135–45.

D. W. Gaskins, Jr., "Alcoa Revisited: The Welfare Implications of a Second Hand Market," *J. Econ. Theory*, Mar. 1974, *7*, 254–71.

R. Latham, "Quantity Constrained Demand Functions," *Econometrica*, Mar. 1980, *48*, 307–13.

M. Okuno, A. Postlewaite, and J. Roberts, "Oligopoly and Competition in Large Markets," *Amer. Econ. Rev.*, Mar. 1980, *70*, 22–31.

L. Singer, "Microeconomics of the Art Market," *J. Cultural Econ.*, June 1978, *2*, 21–39.

H. von Stackleberg, "Price Discrimination in an Arbitrarily Divided Market," *Int. Econ. Papers*, No. 8, 1958, *8*, 65–73.

P. L. Swan, "Optimum Durability, Second-Hand Markets, and Planned Obsolescence," *J. Polit. Econ.*, May/June 1972, *80*, 575–85.

______, "Alcoa: The Influence of Recycling on Monopoly Power," *J. Polit. Econ.*, Feb. 1980, *88*, 76–99.

Revenue Implications of Money Creation under Leviathan

By GEOFFREY BRENNAN AND JAMES BUCHANAN*

Most governments possess a monopoly franchise in the creation of money. Economists provide an analytical justification for this institutional arrangement either in terms of the use of monetary aggregates for macro-economic stabilization or in terms of the alleged inability of competitive markets to generate tolerably efficient monetary results. Any complete case for the government's monetary monopoly must, however, depend on a comparison of market and political arrangements, a comparison that requires predictions about how governments are likely to behave once a monopoly franchise is assigned. Similarly, such predictions are crucial in evaluating restrictions that might be imposed on the government's exercise of its money creation power—in designing a "monetary constitution."

In the analysis of political arrangements, the revenue implications of the money creation power are probably more significant than considerations of either macro-economic stability or optimality in the money supply. Although those revenue implications are incidental to demonstrating the nature of market failure in monetary arrangements, they are fundamental in understanding how the government might exploit a monopoly in money creation, once granted.

Our interest in the revenue effects of money creation stems from a broader study of constitutional restrictions on the revenue-raising authority of government (see our book). The power to create money is naturally encompassed in this. Restrictions on the revenue-raising power must embody restrictions on the power to create money; consequently, the *fiscal* constitution has important implications for the monetary constitution. In this paper, we examine both the revenue implications of money creation, and desirable consitutional restrictions on the money creation power within the context of a specific model of political processes.

I. Money Creation as a Revenue Device

Money creation may involve an addition to real government revenue in three distinct ways. First, there is the possibility that any newly created money can be used directly to purchase real goods and services from private citizens. Second, any attendant inflation will reduce the real value of any outstanding government liabilities that are specified in nominal terms, including specifically outstanding government bonds. Thirdly, inflation may interact with a progressive tax rate structure to increase real tax revenues. In this paper, our analysis is focused on the direct revenue from money creation as such. The discussion could be extended to include the effects of money creation on government interest-bearing debt, but we do not attempt such extention here. The analysis does not bear at all on the matter of income tax revenues.

We begin by considering some conceptual "initial" period in which a society converts its pure barter system into a fully monetized one under government aegis. The government creates a stock of money, M, in the form of pieces of paper that can be used as a medium of exchange. Because money provides a service as a facilitator of transactions, individuals will pay for monetary instruments by giving up real goods and services in exchange for it. The total amount

*Center for the Study of Public Choice, Virginia Polytechnic Institute and State University. Support for this research was made available by the National Science Foundation.

of goods and services so relinquished will have, by definition, a real value of M dollars of real goods, at initial period prices.

Suppose now that in the period subsequent to the initial period, government authorities increase the money stock by ΔM. This increase will, whatever its influence on the price level, clearly have some positive real value. Individuals will give up real goods and services to obtain the transaction services of an additional dollar, even if the goods given up per dollar are somewhat less than in the initial period (i.e., even if the price level is somewhat higher). Accordingly, the government has the capacity in any period in which cash commands any value at all to obtain real revenue in that period by appeal to the printing press.

II. Natural Limits Expectations and Retroactivity

With any conventional revenue instrument (such as an excise tax on beer), there are natural limits on the real tax revenue that can be derived. Increasing the rate of tax will, beyond some point, reduce the tax base sufficiently to *reduce* total revenues. The point at which this occurs is exactly analogous to the point of maximum profit for a monopolist, and the maximum tax revenue obtainable is precisely identical to the profit a pure monopolist would derive if granted a monopoly franchise in the sale of the taxed item.

Are there such natural limits in the money creation case?

To answer this question we focus on an important difference between money creation and (most) conventional revenue instruments. With a tax on beer, for example, an increase in the tax rate, *ceteris paribus*, automatically reduces the quantity of beer purchased, because the price of beer necessarily rises. In the money creation case, however, while it is true that future increases in the stock of money will, *ceteris paribus*, increase the cost of holding money, the precise magnitude of those future increases cannot, in general, be known at the time when the decision to hold cash balances is made. There is, therefore, an extra dimension to the money creation case—it is only to the extent that current additions to the money stock influence *expectations* about the future additions to the money stock that there is a connection between the size of the "tax rate" and the "tax base."

Let us return to our "initial" period. The real value of goods and services relinquished in return for the services of money in that period depends both on the demand for the transactions services money provides, and on the expected cost of holding money. For the purposes of this discussion, we assume that money earns no interest as such; hence, the actual cost of holding cash balances in any period is the actual real interest on interest-bearing assets forgone plus any reduction in the value of money that occurs over that period.

In determining the size of the cash balances individuals wish to hold in the "initial" period when money is created, they must form expectations of both future real rates of return and future rates of inflation. Expectations about future rates of inflation depend in turn on expectations about how the government will act in creating new money in the future periods.

There is, of course, one setting in which the distinction we have drawn here between *actual* increments to the money stock and *expected* increments is irrelevant. This is the case in which there is a fully binding, predetermined "monetary constitution," in which the entire future history of the money stock is charted in the initial period. In this situation, all actual increases in the money supply will be anticipated—and if the monetary constitution is binding, all such expectations will be fulfilled. This is essentially the setting analyzed by Martin Bailey. The revenue implications of money creation are identical with those of a tax on some good —with the peculiarity that money is virtually costless to produce. There is a maximum revenue obtainable from money creation which, given costless production, occurs at that level of inflation (or deflation) at which the elasticity of demand for transactions services is unitary.

The absence of a predetermined and binding monetary constitution, however, drives a wedge between *actual* increments to the money stock and expectations of those increments. It is as if the implied tax rate is set by government *after* rather than *before* decisions are taken by individuals as to how much of the taxable base to possess. This retroactivity is a basic feature of money creation. The individual who decides to hold balances in any period does so in the light of *expectations* about the future course of the money supply—expectations which are necessarily formed *before* the government decides how much new money to create in those future periods. In holding cash balances at all, the individual becomes hostage to the good graces of government. He becomes liable to exploitation in a way that is not feasible in the case where "tax rates" are announced *ex ante*.

This is a characteristic that money creation shares with all taxation of wealth, whether privately or publicly created. A current tax on income from capital is, of course, nominally equivalent to some tax on the capital stock; but a tax of more than 100 percent on current income can be avoided by leaving one's capital stock idle in the current period, whereas an equivalent tax on the capital stock cannot be so avoided. There is virtually no scope for escaping the tax. Because a *current* capital tax is a tax on the outcome of decisions made in *previous* periods to save and invest, it is retroactive in the same sense as is the inflation tax (except where the future course of such tax rates is specified *ex ante* in a binding way). Where money balances are different from conventional capital, however, is that the money stock does not diminish physically. Unlike stocks of wine, the asset cannot be drunk: unlike physical machinery, the stock of money cannot be worked at a rate that leads to premature decay. In this sense, the retroactivity embodied in new money creation is more striking and the scope for exploitation of the money creation power more spectacular than for wealth taxes generally.

Public finance specialists, focusing on more traditional revenue instruments, have regarded retroactivity an undesirable characteristic of taxes and/or tax changes. The reasons for such antipathy are not entirely clear: the matter is rarely discussed explicitly, and the retroactivity property of wealth taxes has not, to our knowledge, generally been noted. In what follows, we shall argue that such retroactivity is a "bad thing." Our objective is to establish a case for a monetary constitution, based not on any macro-economic stability features of a fixed money rule but on the predicted rational calculus of an individual at some quasi-constitutional level when he considers alternative outcomes that might emerge from the government's possession of open-ended money creation power. A crucial ingredient in the argument is the model of political process we adopt, and from which predictions about the behavior of government can be made.

III. The Model of Government

The central thrust of public choice theory suggests skepticism about the capacity of majoritarian electoral processes to constrain governments. Notwithstanding periodic elections, considerable discretionary power remains in the hands of public officials—the "agenda setters" of the political process. In our model of politics we abstract from electoral considerations entirely. In one sense, this may seem akin to the "benevolent despot" model of political process that dominates orthodox discussion. However, we reject the presumption that the discretionary power vested in public officials will invariably be exercised in the "public interest." Without denying the possibility of "moral behavior" on the part of those who hold discretionary power, the asymmetry of such an *assumption* with the behavioral assumptions made elsewhere in economics seems methodologically outrageous, as well as highly questionable empirically. For the purposes of comparing political and market institutions, the central question is whether the institutional structure is such as to translate private interest into public interest: the crucial issue is whether the operation of the particular institution serves to generate socially desired outcomes from the interaction of privately motivated agents. This issue is *avoided*, not answered, by the expedient of assuming political agents to be privately motivated solely by a concern for the "public interest." Moreover, no case for *any* form of institutional constraint — electoral or otherwise—on any aspect of government behavior can be made if it is simply assumed that all coercive power is to be exercised benevolently: such constraints could only prevent the saints from doing good! In this sense, the benevolence assumption is irreconcilably at odds with the basic philosophical underpinnings of constitutional government.

Accordingly, we assume a model of government in which political agents *do* exercise discretionary power, and are motivated solely by private interest in doing so. This private interest takes the form of "surplus maximization," where the surplus in question is the excess of total revenue collections over expenditures on public goods production that government is legally/constitutionally obliged to make. It seems plausible to argue that such surplus will increase with revenue. If so, surplus maximization requires the maximization of total revenues collected from any constitutionally assigned fiscal instruments. It is this simple caricature that represents our "Leviathan" model of political process. As an entity, government is taken to maximize revenues from whatever revenue sources are granted to it by constitutional authorization. Although the model is a caricature, it is a useful one, particularly in a constitutional setting, because in large measure constitutional rules should be designed explicitly to deal with "worst possible cases." Furthermore, although we do not advance the Leviathan model primarily as a positive description of governmental behavior, we do believe it to have at least as much descriptive value as the benevolent despot alternative. The hyper-inflationary experience of twentieth-century history indicates that this sort of revenue-maximizing behavior by government is a contingency worth protecting against.

IV. Revenue Maximization under Permanent and Probabilistic Leviathan

Suppose then that the government seeks to maximize revenue. What strategy will it follow in seeking this end, given that it expects to hold political power permanently?

Clearly, to the extent that there is a determinate positive connection between current inflation and current expectations about future inflation, something like the conventional "revenue-maximizing inflation rate" as derived by Bailey might emerge. There will remain, in each period, a short-term potential for monetary exploitation: current money holders run the risk of their cash balances being reduced in real value spectacularly without any possibility of current adjustment. But such a strategy will no doubt affect inflationary expectations drastically, reduce desired money balances in future periods, and reduce the future revenue potential of the money creation power. Despite the continuing short-term scope for maximum inflation, therefore, the permanent Leviathan may forgo such gains and adopt a policy of restraint. In so doing, there may emerge from this strategic interaction between government and money holders an equilibrium not unlike the Bailey revenue maximum. But any stable adjustment between holders of balances and government must remain precarious. If Leviathan acts in accordance with a finite time horizon, the temptation to default on the pre-announced rule, or to depart from its pattern of behavior established in prior periods, increases. As the final or terminal period is approached, Leviathan will find it advantageous to confiscate the capital values of previously held money balances, regardless of its behavior in earlier periods.

The finite time horizon case becomes, of course, the more relevant one in a regime of competing parties in a democracy, where governments in power rotate regularly but with considerable uncertainty as to specific electoral results. In this setting, if we limit ourselves to the revenue objective, there is little or nothing to be gained from a policy of monetary restraint. Faced with uncertain and time-bound electoral constraints, a government, even if it is characterized by very modest revenue-seeking proclivities, will have extremely strong incentives to capture the revenue potential inherent in any existing value of real cash balances held by the public. Equally, even if the individual reckons that government will generally be "benign," he may also reckon that the government may take on Leviathan proclivities occasionally. In such a "probabilistic" Leviathan setting, the possibility of being totally exploited by hyperinflation is one with which the individual holder of cash balances will have to reckon. Indeed, we should note that the citizen may gain virtually nothing from an occasional success of "good" government. Monetary exploitation may be actually greater in a regime of rotating "good" and "bad" governments than it would be under a regime of continuing and permanent Leviathan.

V. The Monetary Constitution

In the situations described above, it will be advantageous for both the prospective holders-users of real money balances and the government (even if the latter is described accurately as a permanent Leviathan) to agree on a genuine constitutional rule that will constrain the issue of money along some predictable path. Such a rule would, of course, restrict the revenue-seeking flexibility of Leviathan. But, as the analysis suggests, such a constraint might succeed in generating generalized expectations that money issue will be kept within bounds. In the process, Leviathan may well actually secure *greater* value from its money creation authority than it would if it remained unconstrained. Government may, therefore, agree to an enforcible constitutional rule, even if this rule is accompanied by the establishment and maintenance of an enforcement agent, the judiciary, that will be empowered to ignore direct governmental controls (see William Landes and Richard Posner).

We can, therefore, offer an explanation for the emergence of constitutional monetary rules, even under Leviathan government. But the more general constitutional question concerns the initial delegation of money-creation authority to government. As the analysis implies, it seems almost inconceivable that *open-ended* delegation of money-issue authority to government would emerge from a rationally based constitutional calculus.

Whether or not a specifically limited power of continuous money issue would be granted to government is more debatable. One certainly cannot preclude the possibility that some (fixed) positive inflation rate may be fiscally desirable under certain institutional arrangements. What one can rule out is the possibility that no constraints at all will be placed on the government's exercise of the money creation power. The case for some monetary constitution, under any remotely plausible model of government behavior, seems almost unassailable.

REFERENCES

M. J. Bailey, "The Welfare Cost of Inflationary Finance," *J. Polit. Econ.*, Apr. 1956, *64*, 93–110.

Geoffrey Brennan and James M. Buchanan, *The Power To Tax*, New York: Cambridge University Press 1980.

W. Landes and R. Posner, "The Independent Judiciary in an Interest-Group Perspective," *J. Law Econ.*, Dec. 1975, *18*, 875–902.

Inflation, Bank Profits, and Government Seigniorage

By JEREMY J. SIEGEL*

Analysis of revenue from government monetary expansion, referred to as seigniorage, frequently postulates an aggregate demand function for government fiat money. Rarely is it recognized in these formulations that the demand for fiat money is composed of two distinct components:[1] a direct demand for currency and an indirect demand for bank reserves derived from the public's demand for deposits and government reserve requirements. It is the purpose of this paper to explore how bank regulation and the state of competition in the banking industry affect both government revenue and the bank profits.

I. Government Seigniorage: The Basic Model

The model consists of a three-asset economy: government fiat (high-powered) money, with zero yield which serves as currency and reserves of banks; deposits issued by the banking system, and real assets, yielding a fixed real rate of return, set at zero for convenience. There is no uncertainty, the economy experiences no real growth, and the government controls the nominal supply of high-powered money and the (binding) reserve ratio on deposits issued by banks. Under these conditions the long-run equilibrium is marked by an inflation rate which is equal to rate of growth of high-powered money and is equivalent to the market rate of interest.

In this long-run equilibrium, government seigniorage can be expressed algebraically,

$$(1)\quad S=\dot{H}^n/P=(\dot{H}^n/H^n)(H^n/P)$$
$$=\pi H(\pi,r)=\pi C(\pi,r)+\pi k\,D(\pi,r)$$

where S is real seigniorage per unit time, H^n is the nominal supply of high-powered money, P is the price level, π is the rate of inflation, and H is the real demand for high-powered money which in turn is composed of real currency demand C plus the reserve ratio k times the real deposit demand D. Currency and deposit demands are a function of all alternative rates of return; in particular, the market rate of interest (inflation rate) π on real assets and the bank deposit rate r, so that $C_\pi, C_r \leqslant 0, D_\pi < 0, D_r > 0$.

For a given reserve ratio, seigniorage is maximized when $dS/d\pi$ is set equal to zero. This yields the standard result that the total elasticity of demand for high-powered money with respect to the inflation rate be set at unity. If, as the empirical literature suggests (see Cagan, 1956, and Jacob Frenkel), the elasticity of demand for government money is an increasing function of the inflation rate, then seigniorage increases with inflation to the point of unity elasticity and then falls. In order to investigate how the structure of the banking system affects the demand for high-powered money, it is necessary to examine the effects of inflation on bank profits and the deposit rate.

II. Inflation and the Banking Industry

First, consider a bank which has a monopoly in the provision of deposits and is subject to constant costs, a reserve ratio of k, and no restriction on the rate r it sets on deposits. Assuming there are negligible real resources involved in producing deposits, then the real profits B of the monopoly bank can be represented

$$(2)\qquad B=[(1-k)\pi-r]\,D(\pi,r)$$

where $(1-k)\pi$ is the average rate of return on its assets (loans and reserves) and r is the

*Associate professor of finance, Wharton School, University of Pennsylvania. I would like to thank Edward Tower and Phillip Cagan for their helpful comments on an earlier version.

[1] A notable exception is found in Phillip Cagan (1972).

rate paid on its liabilities (deposits). The first-order condition for profit maximization is

$$(3)\quad dB/dr = [(1-k)\pi - r]\,D_r - D = 0$$

To determine the effect of inflation on bank profits, the total derivative of (2) with respect to π is calculated as

$$(4)\quad dB/d\pi = [(1-k)\pi - r]\,D_\pi + (1-k)D + (dr/d\pi)[((1-k)\pi - r)\,D_r - D]$$

Since the term multiplying $dr/d\pi$ is zero from (3), it can be seen from (3) and (4) that profits will rise or fall with inflation under the condition

$$(5)\quad dB/d\pi \gtreqless 0 \quad \text{iff} \quad (1-k) \gtreqless -D_\pi/D_r$$

If the demand for deposits is negatively related to their opportunity cost, defined as the difference between the market rate and the deposit rate, (i.e., $D(\pi - r)$, so that $D_\pi = -D_r$), then, for any positive reserve ratio, profits of a monopoly bank must fall. This is due to the fact that the reserve ratio acts as an excise tax on the earning assets of the bank and an increase in inflation is equivalent to raising that tax.

If the characteristics of the demand function for deposits is that the effect of the own rate (D_r) is greater than any cross rate (the dominant diagonal property of asset demands), then $|D_\pi| < D_r$ and the sign of $dB/d\pi$ is ambiguous. If the reserve ratio were near zero, bank profits would rise under inflation since a monopoly bank would experience a rise in deposit demand due to the shift out of currency. However, since there is empirical evidence (see Joanna Frodin and Richard Startz) that the effect of the market rate and deposit rate on deposit demand are of nearly equal magnitude, it is most likely that monopoly bank profits will fall.

III. Effect of Banking Structure on Government Seigniorage

Assume the government imposes a profits tax at rate τ on the profits of the banking system. Total revenue R to the government can be written as

$$(6)\quad R = S + \tau B = \pi C(\pi, r) + k\pi D(\pi, r) + \tau[(1-k)\pi - r]\,D(\pi, r)$$

If the banking system is competitive, so that bank profits are zero, then government revenue is identical to that which would obtain under a 100 percent profits tax on a monopoly bank. This can be seen by noting that if $\tau = 1$, $R = \pi C + (\pi - r)D$ which is identical to $S = \pi C + k\pi D$ when the competitive deposit rate r equals $(1-k)\pi$. Maximizing R when $\tau = 1$ is equivalent to being a monopolist in the provision of both currency and deposits. This implies that, as long as the government has the power to set reserve requirements, it can do no better by controlling both the deposit and currency markets than it would by enforcing competition in the banking sector and setting the appropriate reserve ratio.

The first-order condition for revenue maximization with respect to the reserve ratio is

$$(7)\quad dR/dk = \pi\, dr/dk(C_r + kD_r) + \pi D(1-\tau) = 0$$

since the terms involving dr/dk are zero by (3). This expression demonstrates that as long as $\tau < 1$, then $C_r + k\,D_r > 0$, so that k must be high enough so that the demand for high-powered money is positively related to the deposit rate. This is equivalent to stating that deposits and high-powered money are complementary assets. In general, the higher the tax rate, the lower the equilibrium reserve ratio.

If $\tau = 1$, so that the sum of the monopoly bank profits and government seigniorage is maximized, the first-order condition (7) reduces to $C_r + k\,D_r = 0$. In this case the reserve ratio is set low enough so that the rise in the demand for reserves occasioned by a rise in the deposit rate is exactly offset by the decline in currency demand. In fact, if currency demand is insensitive to the deposit rate, then the optimal reserve ratio for the maximization of joint profits is zero.

TABLE 1—REVENUE-MAXIMIZING PARAMETER VALUES UNDER VARIOUS BANKING STRUCTURES AND TAX RATES[a]

Maximand	π^*	k^*	r^*	S^*	B^*	S^*+B^*
Seigniorage under Competition	$\alpha/2b$	β/α	$\frac{\alpha-\beta}{2b}$	$\frac{\alpha^2+\beta^2}{4b}$	0	$\frac{\alpha^2+\beta^2}{4b}$
Seigniorage under Monopoly	$\alpha/2b$	β/α	$\frac{\alpha-3\beta/2}{2b}$	$\frac{\alpha^2+\beta^2/2}{4b}$	$\frac{\beta^2}{16b}$	$\frac{\alpha^2+(3/4)\beta^2}{4b}$
$S+\tau B$	$\alpha/2b$	$\frac{2\beta(1-\tau)}{\alpha(2-\tau)}$	$\frac{\alpha-\beta(3-2\tau)}{2b(2-\tau)}$	$\frac{\alpha^2+2\beta^2(1-\tau)}{4b(2-\tau)^2}$	$\frac{\beta^2}{4b(2-\tau)^2}$	$\frac{\alpha^2+\beta^2(2(1-\tau)+1)}{4b(2-\tau)^2}$
$S+B$	$\alpha/2b$	0	$\frac{\alpha-\beta}{2b}$	$\frac{\alpha^2}{4b}$	$\frac{\beta^2}{4b}$	$\frac{\alpha^2+\beta^2}{4b}$

Note: π=inflation rate; k=reserve ration on deposits; r=deposit rate; S=seigniorage; B=monopoly bank profits; τ=tax rate on bank profits.

[a]Assume $C=\alpha-b\pi$, $D=\beta-b(\pi-r)$.

The first-order condition with respect to the inflation rate is

$$(8)\quad dR/d\pi=\pi C_\pi+C+k\pi D_\pi +kD+\pi\, dr/d\pi(C_r+kD_r) +\tau((1-k)\pi-r)D_\pi+\tau(1-k)D=0$$

Together with (7), these two equations solve for the revenue maximizing reserve ratio and inflation rate for a given tax rate. if $\tau=1$ then

$$(9)\quad dR/d\pi=\partial R/\partial\pi =\pi C_\pi+C+(\pi-r)D_\pi+D=0$$

and the government may set its reserve-maximizing inflation rate as if the deposit rate were fixed. The reason for this is when the tax rate is 100 percent, the reserve ratio has already been set so that the deposit rate does not affect the sum of government and pretax bank profits.

IV. An Example

Assume the following functional forms for currency and deposit demand:

$$(10)\quad C=\alpha-b\pi$$

$$(11)\quad D=\beta-b(\pi-r)$$

Table 1 records the revenue-maximizing inflation rate, reserve ratio, and the resultant deposit rate, level of government seigniorage, and before-tax bank profits under various assumptions about the tax rate and degree of bank competition. It can be seen the seigniorage is lower if the banking system is a monopoly since the deposit rate is lower than under competition, a proposition derived in the text. Also, as noted earlier, when the tax rate equals unity the maximizing inflation rate, deposit rate, and government revenue are identical to those that exist when the government maximizes seigniorage under a competitive banking structure. It is noted that, as the tax rate rises, the revenue-maximizing reserve ratio falls until it is zero when τ reaches unity.

In this example, the revenue-maximizing inflation rate is unaffected by the tax rate. Although this is not generally the case, it is in this example since the sensitivities of deposit and currency demand to the rate of inflation are identical.

V. Conclusions

This analysis demonstrates that the revenue derived from monetary expansion is critically dependent on the state of regulation and competition in the banking industry. In general, a monopoly bank issuing deposits will suffer profit loss under inflation unless the deposit rate affects deposit demand far more than the market rate. Government seigniorage is higher if the banking system is competitive than if it is monopolized since deposit rates are lower under monopoly banking. If the government is maximizing its own seigniorage and also taxing bank profits, then the revenue-maximizing reserve ratio is generally a declining function of the tax rate. No such implication can be made about the revenue-maximizing inflation rate. When the tax rate equals unity, the revenue-maximizing equilibrium is identical to that which would exist if the banking system were competitive.

REFERENCES

Phillip Cagan, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

______, *The Channels of Monetary Effects on Interest Rates*, New York 1972.

J. A. Frenkel, "The Forward Exchange Rate, Expectations, and the Demand for Money: The German Hyperinflation," *Amer. Econ. Rev., Sept*. 1977, *67*, 653–70.

J. Frodin and R. Startz, "The NOW Account Experiment and the Demand for Money," unpublished manuscript, Univ. Pennsylvania 1980.

Who Should Control the Money Supply?

By Earl A. Thompson*

This paper develops a model determining some motivational peculiarities of an optimal monetary authority, an authority inducing a socially optimal inflation rate. The model initially assumes that discretionary policy is desirable and that all possible authorities have the same information. Section II rationalizes the assumption of the desirability of discretionary policy under rational expectations. The assumption of rational expectations is seen to strengthen rather than weaken, as its proponents are wont to assume, the case for discretionary monetary policy. Section III relaxes the assumption that all possible monetary authorities have the same information and thereby immediately exposes a dilemma unique to problem of finding an optimal monetary authority: Individuals possessing the efficient motivation are so peculiar that it is unlikely to find one that also possesses anything close to the information required to carry out first best discretionary policy. The model serves to rationalize both the unusual motivational characteristics and the poor performance of observed monetary authorities.

The model includes both positive and negative welfare effects of the actual rate of inflation. Positive effects occur, for example, when money creation is a superior form of taxation, or when the corresponding inflation is unexpected and therefore serves to redistribute away from the relatively wealthy, net-monetary-creditor class. Negative effects occur, for example, when real resources are used up in the process of changing prices, say because different sellers raise prices at different times, thereby creating socially wasteful shopping opportunities, the magnitude or frequency of which vary greatly with the actual inflation rate.

At the heart of the model is a perfect information, noncooperative game involving a well-intentioned governmental authority and the public. The solutions to such interactions are typically nonoptimal. (This was pointed out through examples several years ago by James Buchanan and my 1974b paper, and has been recently established in an abstract environment by Finn Kydland and Edward Prescott.) My contribution here, as elsewhere (1979, 1980), is 1) to show that certain governmental forms eliminate such inefficiencies without relying on inflexible legislative commitments, and 2) to point out that such governmental forms are the dominant forms observed in the real world.

I. The Optimal Monetary Authority

First, consider a monetary authority with a personal utility function identical to a social welfare function, a function depending, among other things, on the actual rate of inflation I, and, in a strictly positive fashion, on transfers from rich to poor, T. The monetary authority can control only the supply of money at some future date, a variable affecting only the rate of inflation I. The higher the excess of this rate over $E(I)$, the expected rate of inflation, the greater the transfer from creditors to debtors and therefore rich to poor. (Striking evidence that the poor are debtors to the wealthy is found in George Bach and James Stephenson, Tab. 4.)

Describing the relevant, differentiable, welfare-utility function of the monetary authority as $U(T(I-E(I)), I)$, where $U_T > 0, T' > 0$, and $T(0)=0$, the policymaker's optimum I^* satisfies the marginal condition:

$$(1) \qquad U_T T' + U_I = 0$$

This equation describes, of course, an "equity-efficiency tradeoff," a situation in which the marginal efficiency loss from unexpected inflation $-U_I$ equals the marginal

*University of California-Los Angeles.

gain in social welfare from the transfers resulting from unexpected inflation, $U_T T'$.

The expected rate of inflation, which is beyond the control of the monetary authority, has yet to be determined. Assuming rational expectations on the part of the borrowers and lenders, $E(I)=I^*$. The public sees who the monetary authority is, correctly anticipates his solution, and thereby forces a solution value of $E(I)$ to be such that $T^*=0$. This holds regardless of social arrangements. Therefore, with social welfare maximized recognizing the inability to independently vary $E(I)$, the relevant constraint being $T=0$, the constrained maximum occurs at I^{**}, where

$$U_I = 0 \tag{2}$$

The contrast between the normal bureaucrat's optimum and the social optimum is shown in Figure 1, where it is also shown that $I^{**} < I^*$. That is, the socially optimal rate of inflation is strictly less than the normal bureaucrat's optimum. Our normal, redistribution-oriented bureaucrat cannot achieve the optimum for he cannot precommit himself to a certain value of I. If the value of I were down at I^{**}, he could not resist the temptation to inflate. After all, the marginal efficiency loss from slightly more inflation at I^{**} is essentially zero while the marginal redistributional gains are significantly positive. The rate of inflation under this authority must rise to where the redistributional gains from additional inflation are offset by the efficiency losses. At his solution, the authority appreciates the significant efficiency gain from a lower inflation rate, but cannot tolerate the redistribution toward the wealthy that the lower rate would induce.

Now consider a possible monetary authority with a somewhat peculiar utility function in that he has no preference for redistributions to the poor. If he were the authority, I would be set so that $U_I=0$. The public, seeing this, would set $E(I)$ equal to his solution I and the solution would therefore coincide with the true welfare-maximizing solution described in (2), the social optimum at I^{**}.

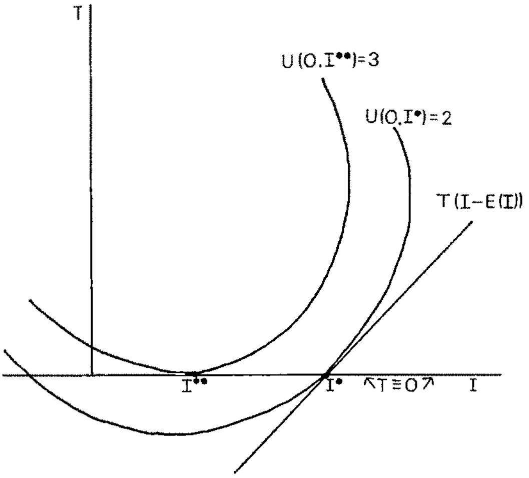


FIGURE 1

When revenue-raising costs are present, there is an additional social benefit of $I-E(I)$ in that it represents a costless lump sum tax. An optimal monetary authority must then have a strictly positive utility for redistribution toward the wealthy in order to keep the authority from the temptation of collecting an unexpected inflation tax. An authority of this sort must himself have sufficient creditor interests. In either case, I call the optimal authority "distributionally neutral" because he always feels a zero net benefit from the unexpected component of inflation.

Most generally, let social welfare depend on both $E(I)$ and I, where the unconstrained welfare maximum $(I^0, E^0(I))$ has $E^0(I) \neq I^0$. See Figure 2; it is assumed that $\partial U/\partial E(I) < 0$ (see Martin Bailey) except at very low $E(I)$ (see my 1977 paper). Since this solution is in fact infeasible, social welfare must be maximized subject to the constraint that $E(I)$ is equal to I. At the resulting optimum $(I^{**}, E^{**}(I))$, increases in I come at the cost of equal increases in $E(I)$ so that the social value of increasing I by an extra 1 percent equals the social cost of increasing $E(I)$ by 1 percent. But a monetary authority with representative preferences will, taking the expected inflation rate as given, expand I to I^*, where its marginal value is zero (Figure 2). An *optimal* authority acts as if he believes it to be an

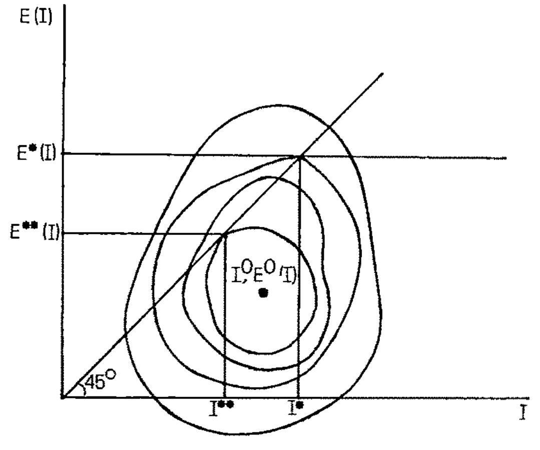


Figure 2

empirical fact that whenever he changes the actual rate of inflation, the expected rate changes by the same amount. Since any authority believing that his actual money supply change is concurrently anticipated regardless of the change he selects would be decidely paranoid, this characterization does not at first glance appear very helpful. But think of our problem as occurring in the first among several periods and specify the affected rates to be the *future*—not current—expected inflation rates. The belief then becomes at least qualitatively reasonable if we also relax the assumption that the public has perfect information over the preferences of the authorities and therefore admit some adaptive expectations. But our authority is extreme; he believes that popular inflationary expectations are perfectly extrapolative in that the current inflation rate his policy produces will be the expected inflation rate for all of the *future* periods.[1] When this authority is also distributionally neutral, he feels relatively insignificant welfare effects of differences between *current*, actual, and expected inflation rates, and therefore feels that the only significant welfare effect of lowering the actual inflation rate is to lower the expected future inflation rate by about the same amount. He will therefore choose an inflation rate that is approximately equal to the optimum at I^{**}, which is less than I^* as long as $\partial U/\partial E(I)<0$ at $(I^*, E^*(I))$.

While the generalized model adds a heavy additional burden on the search for an optimal authority, the model only applies in modern, noncompetitive money economies. In perfectly competitive money economies (see my 1974a, 1977 papers), where competitive interest is in effect paid on all monies, $E(I)$ would have no effect on social welfare other than through its effect on $I-E(I)$, and our less-burdensome initial model would suffice.

The generalized model extends to some degree to any bureaucrat and serves to explain the seemingly excessive concern with "precedents" in successful bureaucracies. But the special model, and the corresponding argument against a redistribution-oriented authority, loses its force for ordinary governmental bureaucrats because of the relatively insignificant potential for unexpected changes in their policy variables to improve the social distribution of wealth.

II. Why Discretion?

An apparent substitute for hiring a monetary authority with motivational peculiarities is to legislatively freeze the money supply or tie it to some objective economic index. But we do not observe such mechanisms. This is presumably because appropriately timed discretionary variations in the money supply provide theoretically efficient offsets to shocks in aggregate demand or supply that are not isomorphic to changes in any objective economic indicator.

Such offsets are socially valuable as they prevent significant, inefficient variations in production and employment due to overly inelastic wage expectations of certain kinds of labor, their faulty expectations being based on their inability to distinguish a shock altering the wage level from one altering only relative wages between different locations or occupations. An advantage of this "confusion" theory of unemployment is that it enables us to model Keynesian-type unemployment within a fully competitive

[1] Recent empirical estimates indicate that a 1 percent increase in the actual inflation rate increases the popularly expected rate by less than 0.1 percent (see Rodney Jacobs and Robert Jones).

temporary-equilibrium model (see my 1974a, 1977 papers) rather than naively imposing a question-begging fixed price on a standard competitive model (see John Hicks). For some reason, our temporary equilibrium theorists have failed to see this and have insisted on modeling Keynesian unemployment by imposing an artificial money wage on an otherwise unconstrained temporary-equilibrium model. These authors have, en masse, failed to see that inefficient unemployment in a competitive environment is the sensible result of certain laborer's having an incorrect perception of a *future* wage rate (for a job in a new location or occupation) and therefore arises in a simple unconstrained temporary equilibrium. (For a survey of the numerous, inappropriate, unemployment models built upon the original oversight of Hicks, see J. M. Grandmont or Allen Drazen.) The importance of this error is that it has obscured the powerful policy guide provided by a temporary equilibrium theory of unemployment. In particular, the theory implies that the inefficiency arising from involuntary unemployment is due solely to inaccurate future wage expectations. For example, if we were in the depths of a depression with an extremely high level of unemployment, the unemployed workers, having discovered their past errors and understanding that the money wage path is lower than they had thought, would return to employment along an optimal time path without any interventionist monetary policy; effective expansionary policy would fool them into returning to work too quickly! Policy intervention is justified only during a time period in which the policymaker knows the future wage level better than the uninformed laborers. In such a period, an optimal policy is to change the money supply to fulfill the wage expectations of the uninformed laborers. (The policy does not induce errors by the informed workers as long as they have rational expectations and therefore see that the exogenous shock has no effect on the postpolicy money wage level.) Optimal policy is preventive, not reactionary; we cannot help by "curing" a depression, we can only help by preventing one. And we could, if we had perfect information regarding near-future equilibrium wage levels, have perfect, completely preventive, countercyclical monetary policy.[2]

With perfect, preventive policy, the Section I solutions, all of which have the property that $I-E(I)=0$, would also represent optimal discretionary policy solutions regardless of the type of monetary authority. Only here the value of I is adjusted for unsystematic aggregate shocks so as to justify the value of $E(I)$ set at the nonshock solution value of I to the policymaker's nonshock optimization problem. Exogenous aggregative shocks are neutralized; they affect neither I^* nor $E^*(I)$. Those satisfied with perfect policy models can proceed to Section III.

Now suppose that the unsystematic changes in aggregate demand cannot be perfectly observed by the monetary authority. $E(I)$ will differ for different individuals, the average $E(I)$ exceeding I when there is "involuntary unemployment" and falling short of I when there is "cyclical overemployment." And only the mathematical expectation of the average $E(I)-I$ will, under rational expectations, be set equal to 0. The variance σ of the distribution of $I-E(I)$ about its zero mean will now negatively enter our optimal monetary authority's utility function, as will any current absolute deviation of I from $E(I)$. The $I-E(I)$ is, under rational expectations, determined by off-trend changes in the money supply M, and other labor-confusing aggregating demand shocks A. Hence, $I=E(I)+g(M,A)$. Since the authority cannot perfectly observe A, M cannot be perfectly correlated with A so as to neutralize its effect and make $g=0$ always. If a particular aggregate shock is unobserved by the monetary authority, the deviation of I from $E(I)$ is unavoidable.

[2] While our temporary equilibrium-policy theory need not be restricted to a single commodity, labor is the only competitively marketed commodity we know of for which government authorities often have systematic information advantages concerning future market prices over many of the market transactors. The absence of a capital market for human capital prevents specialist-speculators from entering and revealing variations in labor values in different areas to the less-informed owners of labor.

But if the shock is observed, the authority sets M so that $g=0$ and no real fluctuation occurs. Because $I=E(I)$ under these conditions, there is no change in the Section I solution. As such a policy also reduces σ to a minimum given $E(I)$, there is again no conflict with the policy framework of Section I despite the authority's now-imperfect ability to observe labor-confusing aggregative shocks.[3]

We have escaped the familiar short-run policy tradeoff between unemployment and wage inflation in that our optimal, strictly preventative, wage increases are increases from otherwise subnormal to normal optimal wage levels. A policy producing unexpected wage changes is *never* desirable!

Moreover, the assumption of rational expectations prevents the policymaker from introducing economic fluctuations by imposing a pointess "*systematic*" policy function, a function of popularly observed variables, on the economy. Suppose, for example, that he systematically expands the money supply in the depths of a depression. This would introduce equally greater values of both I and $E(I)$ without affecting the optimality of the *laissez-faire* recovery path. Rational expectations protects us against ill-conceived systematic countercyclical policies, and therefore provides new support for discretionary monetary policy. The popular impression to the contrary is based on the above-described misformulation of the underlying welfare economics of labor-confusing aggregative shocks.

III. Empirical Application

Summarizing the above argument, our optimal monetary authority: 1) has distributionally neutral, politically independent, preferences; 2) erroneously believes that increases in the actual rate of inflation increase expected rates of inflation by nearly equal amounts, and 3) is both willing and able to respond to aggregative demand or supply shocks with monetary shocks before the former have significant, popularly observable, effects.

Someone satisfying the first two rather demanding characteristics is unlikely to satisfy the third. A distributional conservative with grossly exaggerated views of the sensitivity of future-to-current inflation rates is unlikely to be either—let alone both—willing or able to regularly intervene to neutralize aggregative demand or supply shocks before they substantially affect the economy. Since most modern monetary authorities apparently do come close to satisfying the first two criteria, they should not be expected to also be well-informed interventionists. Thus, it is not surprising, for example, that *U.S.* monetary authorities during the 1970's allowed unjustifiable, duplicate booms by failing to tighten money supplies in the face of obvious shifts down in the world demand for dollars (1973, 1978) and allowed painful, duplicate recessions by failing to loosen money supplies in the face of obvious jumps in nonlabor input costs (1974, 1979). Furthermore, to the extent that we have rationally foregone some of the first two attributes to obtain less-incompetent fine tuners, the model rationalizes our recently observed tendency to produce somewhat excessive inflation rates.

REFERENCES

G. L. Bach and J. B. Stephenson, "Inflation and the Redistribution of Wealth," *Rev. Econ. Statist.*, Feb. 1974, *56*, 1–13.

[3] The minimization of σ does not imply the minimization of the variance of output. The prevention of some output fluctuations increases the magnitude of others. Once workers figure out that fewer changes in the aggregate wage level actually occur, they are more likely to rationally mistake any given aggregative wage change for a relative wage change. As a result, employment will be more susceptible to an unneutralized aggregative shock. Nevertheless, there is still an improved allocation of resources even if the sum of the output variations is unaffected. This is because workers are simultaneously adjusting better to the relative wage shifts that must also be occuring in order to keep expectations rational. That is, a relative wage decrease is simultaneously more likely to be interpreted as such and correctly adjusted to with a job switch. So the "natural," or "frictional," level of employment expands and becomes more efficient as the workers rationally adjust to the sometimes-effective countercyclical policy (see my 1977 paper).

M. Bailey, "The Welfare Costs of Inflationary Finance," *J. Polit. Econ.*, Apr. 1956, *64*, 93–110.

J. M. Buchanan, "The Samaritan's Dilemma," in Edmund S. Phelps, ed., *Altrism, Morality, and Economic Theory*, New York 1975.

A. Drazen, "Recent Development in Macroeconomic Disequilibrium Theory," *Econometrica*, Mar. 1980, *48*, 283–307.

J. M. Grandmont, "Temporary General Equilibrium Theory," *Econometrica*, No. 3, 1977, *45*, 535–73.

John R. Hicks, *Capital and Growth*, London 1968, chs. 3–8.

R. L. Jacobs and R. A. Jones, "Price Expectations in the United States, 1947–45," *Amer. Econ. Rev.*, June 1980, *70*, 269–77.

F. E. Kyland and E. C. Prescott, "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *J. Polit. Econ.*, June 1977, *85*, 473–92.

E. A. Thompson, (1974a) "The Theory of Money and Income Consistent with Orthodox Value Theory," in George Horwick and Paul A. Samuelson, eds., *Trade, Stability, and Macroeconomics: Essays in Honor of Lloyd Melzer*, New York; London 1974.

______, (1974b) "Taxation and National Defense," *J. Polit. Econ.*, July/Aug. 1974, *82*, 755–81.

______, "A Reformulation of Macroeconomic Theory," work. paper no. 91, Univ. California-Los Angeles May 1977.

______, "An Economic Basis for the National Defense Argument for Aiding Certain Industries," *J. Polit. Econ.*, Feb. 1979, *87*, 1–36.

______, "Charity and Nonprofit Institutions," in Kenneth Clarkson and Donald Martin, eds., *The Economics of Non-Proprietary Institutions*, Greenwich 1980.

The Inflation Process: Where Conventional Theory Falters

By JAMES W. DEAN*

This paper suggests that conventional economic theory has tended to ignore the *inflation transmission process*, and sketches some of the reasons why that is so. Harvey Leibenstein's paper in this session is intended as a companion piece; it outlines an unconventional theoretical approach to the inflation process.[1]

Inflation just means rising prices. Conventionally, however, economists mean something more. Robert Solow, for example, has defined it as "a substantial, sustained increase in the general level of prices" (p. 31). But when inflation is *sustained* at a steady rate for long enough to be anticipated, those affected adversely begin protecting themselves against its consequences. Moreover, *general* inflation implies an absence of relative price changes and their consequent allocative and distributional effects. As economists we agree that inflation has no real consequences once sufficiently long-lived and general,[2] yet we definitionally preclude short-run and relative price changes from the purview of inflation theory!

For present purposes, the transmission stage is defined as inflation before it is universally anticipated and compensated for. Only in its transmission stage is inflation of practical consequence to anyone, but extant theory has little systematic to say. The one aspect of the transmission process that economists do try to analyze is the macroeconomic *price/output* problem: how is a macro shock to demand or supply divided over time between price change and output change? This problem is far from solved, and is perhaps the central bugbear of modern macroeconomics.

Underlying our ignorance of the inflation transmission process are inadequacies inherent in microeconomic theory. Section I briefly discusses some of these inadequacies. Section II illustrates them by looking at conventional treatments of the price/output problem. For space reasons much of what follows consists of bald assertions, and most references have been deleted. The reader is referred for elaboration to the manuscripts listed at the end.

I. Why Micro-Economic Theory is Inadequate to Analyze Inflation

General equilibrium micro theory in the Walrasian tradition cannot illuminate the transmission process because partial equilibria and disequilibria are outside its orbit. The crucial period of gains and losses is suppressed by definition. Walrasian systems are further limited by their concern with exchange but not production phenomena.

General equilibrium models, moreover, assign money a neutral role with respect to relative prices. Though Walrasian systems fix relative prices, they cannot fix absolute prices. The latter requires a separate monetary model. The resulting "classical dichotomy" stunted economists' ability to analyze the price/output problem for decades. The dichotomy also rules out inflation as the result of relative price changes; even today many economists confuse Walrasian systems with reality by suggesting that oil price hikes cannot cause inflation.

*Simon Fraser University and Columbia University Graduate School of Business.

[1] Both papers are based on our forthcoming book.

[2] Even the "welfare cost" of inflation—much analyzed but of arguable empirical importance—exists only because one price, the interest rate on money, is prevented from rising.

How money enters the economy, and who gets it first, cannot be addressed in a Walrasian framework. Money must enter exogenously, as from a helicopter, and the disequilibrium process of picking it up and passing it around until desired cash balances are realized, must be elided. Thus monetary theorists rarely discuss the endogenous, uneven process whereby borrowers' demand for bank loans become money, or the pattern whereby government balances are released into private hands. Our much-lauded theory of money demand lacks micro-economic foundations, and our theory of money supply is deficient insofar as the money supply responds to demands for it.

Partial equilibrium micro theory, in the Marshallian tradition, provides well-developed static models of the production and exchange activities of maximizing households and firms under perfect competition and under perfect monopoly. Most economic activity, unfortunately, takes place under imperfect competition and oligopoly. This would not matter were the real world readily represented by hybrid models, built from borrowed bits of the two ideal types. But it is not. Moreover there is no evidence that firms and households universally maximize. Let us consider, in turn, oligopolies, imperfect markets, and the maximization hypothesis.

Oligopoly firms have the special characteristic that their actions are interdependent. If one oligopoly raises its price, others may or may not follow. Conventional theory cannot handle the resulting indeterminancy. Game theory handles oligopolies more readily, but does not plug neatly into standard microeconomics.

Standard theory states, for example, that (for given cost curves) prices are more responsive to demand fluctuations under perfect competition than under pure monopoly. Yet oligopoly pricing behavior is not always better described by a competition model as the number of firms increases, or by a monopoly model as the number of firms decreases. Evidence by David Qualls shows that over a range, oligopolized industries with *fewer* firms change their prices *more* markedly over the business cycle. More firms apparently reduces possibilities for collusion aimed at varying prices sufficiently to maximize monopoly profits over the cycle, and leads instead to price rigidity for fear of a negative-sum price war.

Imperfect markets arise because relationships are personal or otherwise ideosyncratic, because institutions, governments and customs impose constraints on maximization and erect barriers preventing arbitrage, and because information is incomplete. Over the last two decades, "information theory" has yielded the important result that with imperfect information, otherwise competitive firms acquire monopoly power.

Conversely, firms with monopoly power require information not needed by competitive firms (see Kenneth Arrow). Under perfect competition, the firm needs no information about its customers' demand schedules; it simply charges the going market price. But if firms with monopoly power are to price optimally, they need demand information that is rather difficult to come by. They need to know the schedule of quantities that *would be* demanded over the range of feasible prices, a schedule that is purely hypothetical. They may undertake market surveys, but they can never really know the demand schedule they might face except after potentially costly experimentation with alternative prices. Though the profit-maximizing price may be arrived at eventually, it is likely approached via protracted Darwinian routes. Firms with monopoly power therefore resort to rules of thumb, like cost-plus pricing, which leads to significant price differences for the same product.

"Monopoly power" is a troublesome concept in and of itself. As used above it means the ability to move *along* downward-sloping demand curves, and optimal price is readily fixed in theory by the conventional static model of the profit-maximizing monopoly. But firms can also *shift* the demand curves they face, via advertising and the like. One might distinguish between "static" and "dynamic" monopoly power; the latter is outside the purview of our standard theory. Yet dynamic market power in the labor market, where unions seek to shift the demand for workers' services so that wages rise without reducing employment, is the essence of wage push inflation.

Whenever a seller with dynamic monopoly power faces a buyer with dynamic monopsony power, the possibility of *bargaining* arises. Conventional theory cannot even in principle fix a determinant price. Potential interaction between sellers themselves or buyers themselves complicates the outcome still further.

Maximization. Micro theory's most basic assumption is that economic agents universally maximize. Testing the assumption for individuals is tricky since their maximand, "utility," is unobservable. Many observable phenomena, like downward-sloping demand curves, are not only consistent with utility maximization, but also with a host of other decision rules (see Gary Becker). But discriminating laboratory tests are possible and they do not support the maximizing assumption. Apparently people fail to behave even *as if* they had undertaken maximizing calculations (see Herbert Simon).

The presumed maximand for firms *is* observable, and evidence on profit maximization is mixed at best. Case studies find that intra-organizational decision making rarely centers on profit maximization for the firm as a whole. Therefore a few economists have proposed theories of the firm involving alternative maximands, transactional costs, evolution, learning, or limited rationally.

A related limitation of conventional micro theory is that it typically defines households and firms as atomistic units rather than as collections of interacting individuals. Within the household, for example, spouses, parents and children take on agent/principal roles, as do managers and owners within the firm. Agents may or may not act exclusively for their principals.

Nearly half our national income is spent for us by agents whom we, as principals, elect to government. Yet no theory of government has been blended into the mainstream of economic theory. Since government is a monopoly supplier of most of its services, and since demand elasticity for many of these services is very low, it can potentially price rather arbitrarily. This potential is heightened by the separation of supplier and user costs inherent in public goods and, relatedly, the coercive nature of taxation. Thus governments contribute to higher prices and to inflation in two ways: via the prices and wages they pay, and via the price and income effects of the taxes they impose.

We similarly have no theory of public utilities, which are often monopolies, and which may or may not be publicly regulated. Much nominally private sector production, moreover, is heavily subsidized by the state or relies heavily on government contracting: defense production is the prime example. Competition between bids is often illusory or nonexistent.

Supplier and user costs are also sharply separated in the provision of insured services such as automobile repairs and injuries, or medical care. The constraints on raised prices are accordingly weakened. As the demand for these services seems presently to be income elastic, the problem has been exacerbated with economic growth.

II. The Price/Output Problem

A fundamental aspect of the inflationary transmission process is the division over time of an initial demand or supply shock between price and output changes. This section illustrates some of the above inadequacies of our micro-economic tools by sketching three models of short-run pricing that bear on this problem: models of administered pricing, cost-plus pricing, and expectations-based pricing.

Administered pricing means price stickiness as demand fluctuates. Evidence shows that concentrated and manufacturing industries administer prices more than do competitive and primary industries.

To what extent can conventional theory rationalize this evidence? The monopoly firm's profit-maximizing price is $MC\cdot\varepsilon/(\varepsilon-1)$, where MC is marginal cost and ε is demand elasticity. With ε constant, price will move strictly in proportion to changes in MC. And since for the perfectly competitive firm price equals MC, it too should change price in proportion to changes in *MC*.

If, however, competitive supply and therefore *MC* becomes inelastic as output rises, the competitive price converges to the monopoly price. As the demand drops, encountering more elastic *MC*, the monopolist, in contrast to the seller facing competition, can restrict supply and hold his price as far above *MC* as demand inelasticity permits. The seller facing competition cannot "administer" price in this way. His selling price will drop sharply in business cycle downturns, along with *MC*.

Thus price stability, in the conventional model, depends on two factors: (1) the rate at which marginal costs rise as output rises: and (2) the degree of collusion to restrict production or sales. Prices of primary products are therefore stable only to the extent that fewness of firms permits monopoly pricing. Prices of industrial commodities, whose variable costs are typically horizontal over a wider range, tend to be relatively rigid even when the industry is competitive.

There are three reasons why this conventional explanation of administered pricing is incomplete. Each relates to the discussion in Section I. First, firms may not always be maximizers, or at least their maximand may not be profits. Second, the degree of collusion between firms, critical to the above discussion, is left indeterminant. Third, firms may stabilize prices in order to mitigate uncertainty, trading off maximum profits in the process.

Cost-plus pricing (*cpp*) models posit that prices are marked up over average variable cost, *AVC*, according to some simple relatively invariant rule. Over a range with *AVC* constant, $AVC=MC$, and with ε constant the (constant) markup multiplier *could be* $\varepsilon/(\varepsilon-1)$. That is, *cpp* is not necessarily inconsistent with profit-maximum pricing (*pmp*).

However, firms do not necessarily maximize profits. They really do mark up over unit costs, but they rarely price *ex ante* where marginal revenue equals marginal cost. Indeed, they typically lack the information to do so. This would not trouble methodological positivists did markups nevertheless typically equal $\varepsilon/(\varepsilon-1)$ *ex post*. But they do not. Case study evidence frequently contradicts profit maximation. Moreover, percentage markups tend to remain fixed on *AVC* but not *MC* when *AVC* varies; this is inconsistent with *pmp* unless ε is assumed to vary in an extremely contrived fashion.

Of course the *cpp* hypothesis does not come by its behavioral realism costlessly. By abandoning profit maximization it abandons determinancy, since the size of the markup is arbitrary. Worse, the size of the markup over average costs may vary capriciously according to the measure of "unit costs" on which it is based: a fixed percentage markup on any subset of costs leads to a steadily increasing markup on total costs if the subset rises faster than the total.

Cost-plus pricing is consistent with administered pricing to the extent that unit cost, or at least the measure of it on which the markup is applied, is constant over the demand cycle. Unit costs of industrial goods are relatively stable cyclically. Reinforcing this is the usual practice of marking up price over "normal" unit cost, a measure specifically designed to be cyclically insensitive.

But *cpp* does not answer Qualls' evidence on price inflexibility cited in Section I any better than does *pmp*, since it also ignores the potential negative-sum consequences of not colluding. Neither does it address price rigidity to mitigate uncertainty. *Pmp* is more readily adapted to incorporate uncertainty: modern contracts theory, for example, attempts to link wage rigidity to maximizing behavior via risk aversion. But it would be somewhat farfetched to treat most nonwage prices as contracts, particularly since the buyer often faces a take-it-or-leave-it choice. And finally, price and wage setting is strongly influenced by social convention and other inertial elements. These are better analyzed outside the conventional paradigm of continuous universal maximation.

Expectations. Axel Leijonhufvud's 1968 reinterpretation of Keynes' recast price stickiness in terms of Hicksian-inelastic price expectations: Because they expect prices to return to "normal," economic agents who experience increases in the prices of what they buy will not adjust the prices of what they sell. Simultaneously, the neoclassical camp pioneered a class of natural rate models in which wage adjustment to higher prices is incomplete until price increases are perceived as permanent and/or spatially general (see Milton Friedman; Edmund Phelps). The Friedman and Phelps models in a sense go beyond Leijonhufvud's by allowing expectations ultimately to become unit elastic as agents learn inflation's temporal or spatial extent.

But the difference in emphasis is perhaps best understood in view of the contrasting origins the two classes of models ascribe to price shocks. The typical Keynesian shock originates in the real sector and is temporary. The typical monetarist shock originates with money and is permanent. Therefore inelastic expectations are likely to be correct in a Keynesian world, and unit-elastic expectations are likely to be correct in a monetarist world. Second generation, "rational" expectations models allow for both possibilities, suggesting that temporary shocks or shocks about which there is insufficient current information to infer longevity will result in output adjustment, whereas shocks which are correctly perceived as permanent will lead only to price adjustment.

The importance of these models lies in their disequilibrium properties. They are certainly not realistic (who really believes that output departs from its natural growth path primarily because inflation departs from expected inflation?), but they have provided the profession's first tractable technique for analyzing the price/output problem dynamically.

Whether the technique can usefully be applied to micro-economic inflationary transmission processes is unclear. For despite the original Friedman/Phelps reductionist concern with micro-economic labor market motives, the "macro rational expectations" models that have followed tend to submerge micro-economic distinctions. They also (though more by practice than design) tend to divert attention from expectations adjustment (i.e., learning), and from the distinction between spatial and temporal learning processes. What is needed if rational expectations models are to shed further light on the inflation *process* are complementary models of how information spreads over time and through space, and of how agents learn. The naive assumption that agents immediately react rationally to new information, or even that their behavior universally converges toward rationality, ignores reams of psychological research, as well as decision theory in its entirety.

III. Conclusion

This essay began by lamenting the lack of a micro theory of inflation. There followed a litany of the relevant shortcomings of microeconomics. It was noted that conventional theory hamstrings us because it tends to be premised on continuous, short-run maximization that is often palpably inconsistent with actual pricing behavior.

By no means does this call for wholesale rejection of extant inflation theory. But it does call for the judicious introduction of behavioral premises which are not usual in our discipline. That is the object of Harvey Leibenstein's paper in this section.

REFERENCES

K. J. Arrow, "Toward a Theory of Price Adjustment," in Moses Abramovitz et al., *The Allocation of Economic Resources*, Stanford: Stanford University Press, 1959, 41–51.

G. S. Becker, "Irrational Behavior and Economic Theory," *J. Polit. Econ.*, Feb. 1962, *70*, 1–13.

R. Cherry, P. Clawson, and J. W. Dean, "Micro Foundations of Macro Rational Expectations Models," mimeo., Simon Fraser Univ. 1980.

J. W. Dean, "What's Wrong with Inflation Theory," mimeo., Simon Fraser Univ. 1980.

______ **and Harvey Leibenstein,** *Towards a Micro-Behavioral Theory of Inflation*, forthcoming.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, New York:

Oxford University Press, 1968.

E. S. Phelps, "Phillips Curves, Expectations of Inflation and Optimal Employment over Time," *Economica*, Aug. 1967, *34*, 254–81.

D. Qualls, "Price Stability in Concentrated Industries," *Southern Econ. J.*, Oct. 1975, *42*, 294–98.

H. A. Simon, "Rational Decision Making in Business Organizations," *Amer. Econ. Rev.*, Sept. 1979, *69*, 493–513.

R. Solow, "The Intelligent Citizen's Guide to Inflation," *Public Interest*, Winter 1975, *38*, 30–66.

The Inflation Process: A Micro-Behavioral Analysis

By HARVEY LEIBENSTEIN*

This paper covers some essential ideas of a much larger study by James Dean and myself of the way in which the behavior of economic agents, when aggregated, contributes to the process of inflation. Space limitations dictate a compromise between brevity and minimal completeness. Therefore many of the basic ideas will be stated baldly. Qualifying remarks are omitted.

The conventional macro approach can be subsumed under the slogan of "too much money chasing a given amount of goods." It does not explain how specific absolute prices are set. Further, a theory of money demand creation from a *micro* viewpoint is missing. I shall emphasize the world where real live flesh-and-blood humans set prices, and where their activities influence the demand for money.

The essential scheme is simplicity itself. Inflation is determined by incentives that lead sellers, using decision rules based mostly on *conventional* behavior, to raise absolute prices, which in turn creates an increase in the demand for credit which, in its turn, the banking system accommodates to some degree.

Can we start with a set of reasonable postulates about firms and markets consistent with this type of firm behavior? I believe we can. The postulates I focus on (taken from X-efficiency theory) can be summarized by the following phrases: 1) inert areas, 2) mostly nonmaximizing behavior, 3) incomplete "employment" contracts, 4) imperfect markets (including some bargaining power), and 5) a take-it-or-leave-it bargaining style.

I. Basic Postulates[1]

1) *Inert areas*: An inert area is a set of boundaries within which behavior is habitual, routinized, and or conventional. This type of behavior changes when an external stimulus is large enough (and/or of sufficient duration) so that the pressure generated by the consequences of continuing such behavior exceeds certain upper or lower bounds.

2) *The nonmaximization to maximization continuum*: Let us take a procedural view towards decision making.[2] Since most behavior is routinized (inert area postulate), most decision making is passive (i.e., made by *not* making an explicit decision), and possibly nonmaximizing. A stimulus, usually involving *external pressure*, that pierces the bounds of the inert area, is required to *activate* explicit decision making. These involve the choice of decision *procedures* which allow for the consideration of at least some alternative options. The explicit procedure may involve no more than (a) finding a variation in one's routines (i.e., habits or conventional modes of behavior), or it may be more elaborate and involve (b) various degrees of partial or (c) full calculation (i.e., maximization). Generally, the greater the external pressure, the greater the shift in procedure towards fully calculating behavior.

Among decision procedures we especially emphasize choices that result from the search for and use of conventions. These involve coordinating one's behavior with others', or behaving similarly to others: for example, choosing to have one's meals at the same time as others do is a choice determined by convention. Since conventions are solutions to interpersonal coordination problems, they are transindividual.[3] Also note that conventions allow for a number of alternative "choice solutions" (i.e., conventions)—not all of which are optimal. For present purposes I focus on two types: the effort pat-

*Harvard University.

[1] Most of these postulates and their implications are discussed extensively in my 1976 book and 1979 article.

[2] On procedural decision theories, see Herbert Simon.

[3] There exists a well-worked out theory of conventions based on game theory. See David Lewis.

terns that firm members choose to interpret their jobs, given effort discretion; and the markup pricing formulas firms use to determine prices.

3) *Incomplete firm membership contracts*: Compensation is determined by contracts but the related effort is loosely defined or undefined; hence some effort discretion. The components of effort are the activities, the pace at which each is carried out, their quality and sequence.

4) *Imperfect markets*: I focus on markets with a variety of imperfections (due to advertising, trade marks, buyers' limited knowledge, frictions, gaps, etc.) Thus sellers have relative bargaining power as compared with buyers, based in part on market imperfections, and on a greater willingness by sellers to forego the benefits of an exchange.

5) *TILI bargaining*: We assume that sellers of products (i.e., firms) use a take-it-or-leave-it price setting (*TILI*) bargaining style rather than negotiated bargaining. The wage bargaining style is considered later.

II. Basic Propositions

The five propositions consistent with the postulates basic to my theory follow:

1) *The elasticity of demand firms face is almost zero for very small price increases.* Inert areas surround purchasing behavior. Hence, households continue to buy the same bundle of goods, up to a certain point, as prices continue to rise. From the seller's viewpoint this means that he will retain customers despite increases in price, and that he will not lose sales despite increases in price, *up to some point*. Up to some point elasticity of demand is zero, and beyond that point it will slowly rise.

2) *In most markets some lose and some gain in the inflation process.* The inert area and bargaining power postulates imply that up to a point people buy the same amount of a product at a higher price. Most contracts are made in nominal terms and are of various durations. Thus some buyers are likely to be losers in some markets because of inflation. Clearly, gains and losses will in part be determined by the degree to which persons are involved in short or long run contracts with respect to income and expenditures. Since people operate in many markets it is to be expected that on balance some will gain and others will lose as a result of inflation.

3) *Economic agents will try to make up (or more than make up) in their strong markets the losses incurred in their weak markets.* Under maximization, agents try to do as well as they possibly can in each market. How a person does in one market is essentially unconnected with how he does in another. However this does not hold under inert area nonmaximization postulates. The inflation stimulus on behavior may be cumulative. At some point (and/or after some time) the inflation stimulus pierces the inert area bounds which surround the budget constraint. Thus, inflation induces people to try to make up in the markets in which they are strong for losses incurred in markets in which they are weak.

4) *Price increases are likely to result in lower productivity levels than otherwise.* Incomplete contracts imply effort discretion. This is partly accommodated for by price discretion in imperfect markets. An inflationary atmosphere creates an environment in which sellers have to worry less about price competition. Thus the price discretionary area is enlarged. Hence, less external pressure leads to less internal pressure by higher-ups on those lower down and therefore, to some degree, a shift of individual effort levels away from the firm's productivity aims. As a result real costs per unit of output rise.

5) *Some marginal markup formulas will be such that cost increases are more than passed through.* It is important to note that *normal* markup pricing conventions can lead to such results. Many conventions are based on formulas that more than cover variable costs so as to contribute towards fixed costs. Using the same formula in periods within which variable costs are rising but in which fixed costs are rising less than proportionately, will lead to the greater than cost pass through effect. As inflation gets high enough its effects pierce the inert areas covering price formulas so that *new* formulas, including those that raise prices, are

considered. In imperfect markets, firms have to test demand elasticity. Under stable prices this can be costly. Firms may passively retaliate, that is, not raise their prices. It is safer to explore price raising possibilities in inflation. Thus firms may shift to formulas that more than pass through cost increases. Note that there is an asymmetric tendency towards price increases rather than decreases since: inert areas imply short-run losses for price declines; firms do not know long-run price elasticities; and there is the fear the price decreases may stimulate retaliation.

III. The Transmission Process

The five propositions suggest the way inflation spreads. Let us start with some external *stimulus* (or stimuli) that raises prices or costs to some agents. The stimulus may be due to import costs (i.e., oil) rising, or a currency devaluation, or tax increases, or a productivity decline, etc. *Some* sellers respond by raising prices because of the way some conventional markup formulas work, or because the stimulus pierces the inert area pricing bounds. Higher prices set by some lead to initial losses to others. Losers in weak markets attempt to make up or more than make up for these losses in their strong markets. A general price rising tendency sets in which results in a partial loss of productivity, and higher costs than otherwise. This may generate even greater price increases. These forces spread through the economy as economic agents in multiple markets, under which buyers in some markets who are sellers in others, create impulses in those other markets, which in turn generates higher prices. At the same time this increases the demand for credit since those who buy at higher prices have to find the finance to do so. Over time, some labor will gradually move away from weaker markets, and losses and gains will slowly change accordingly.

The above assumes that the stimulus was large enough to *generate* an inflationary process. However, some stimuli may be too small, as the inert area concept suggests. Thus, some agents may be net *absorbers* of inflation stimuli, while others are *neutral* (i.e., they exactly pass through higher costs but not more), while still others are net *contributors* to inflation. The inert area concept suggests that for stimuli of low magnitude the inflation absorbers will dominate, *on the average*, and a movement will be set in motion for inflation reduction. However, at some intermediate levels the contributors dominate and the inflation rate increases towards a maximum level, or rises indefinitely. Inflation may be considerably below the fully accommodated level as a result of central bank actions which restrict money supply growth.

IV. An Algebraic Model of the Inflation Process

In the following, I present a sample[4] (albeit partial) algebraic model of the inflation process. It attempts to capture most of my basic ideas. I start with two identities; a micro identity (1a) and a macro identity (1b).

$$m_i p_{t,i} = p_{t+1,i} \tag{1a}$$

$$M_t P_t = P_{t+1} \tag{1b}$$

where m_i is the marginal markup for firm i (i.e., marginal with respect to last year's price), $p_{t,i}$ is the rate of change of the price i in the last period over the period before. M is the *average* (marginal) markup for all sellers, and P_t is the average price *change* for all prices. Equation (1a) indicates how the marginal markup for firm i applied to its price change determines the price change in the next period. Summing p_i of every seller and the quantities produced, we obtain the change in value of all transactions Z: Thus

$$Z_t = \sum_i p_{ti} q_{ti} \tag{2a}$$

$$\frac{p_{ti} q_{ti}}{Z_t} = w_{ti} \quad \text{for all } i \tag{2b}$$

[4] There are a number of possible models that could capture our basic ideas. Hence this is a *sample* of this class of models. It is extremely simple and intended only to be illustrative.

In order to connect the micro identity (1a) to its macro counterpart (1b) we set

$$M_t = \sum_i m_{ti} w_{ti} \tag{2c}$$

where w_{ti} is the share of firm i (or industry i) in Z_t.

Now we want to reflect the idea that price makers take into account events in multiple markets; that is, changes in product price, the cost of inputs, and changes in the price index of all goods. Hence we write

$$m_i = m_{0i} + m_{1i} + m_{2i} \tag{3a}$$

where m_{0i} is the firm's price increase multiplier which reflects its history (and expectations), m_{1i} makes adjustments for factor-cost increases, and m_{2i} adjusts for changes in the general price index. To express the inert area idea we set

$$m_{0i} = 0 \text{ when } 0 < p_i < \bar{p}_i, \tag{3b}$$

$$\text{and} \quad m_{0i} = k > 0 \text{ when } p_i > \bar{p}_i$$

Similarly

$$m_{1i} = 0 \text{ when } \bar{\bar{F}} < F_i \leqslant \bar{F}_i, \tag{3c}$$

$$\text{and} \quad m_{1i} > 0 \text{ when } F_i > \bar{F}$$

$$m_{2i} = 0 \text{ when } \left[\frac{\bar{\bar{P}}}{p_i}\right] < \frac{P}{p_i} \leqslant \left(\frac{\bar{P}}{p_i}\right), \tag{3d}$$

$$\text{and} \quad m_{2i} > 0 \text{ when } \frac{P}{p_i} > \left(\frac{\bar{P}}{p_i}\right)$$

where F_i is the increase in factor costs between any two periods, where $\bar{F}$ is the upper bound and $\bar{\bar{F}}$ is the lower bound of the inert area surrounding F. Similarly $(\bar{P}/p_i)$ and $(\bar{\bar{P}}/p_i)$ are the upper and lower bounds of the inert area. that surround P/p_i. This means that the price multiplier m_i in equation (1a) depends on the factor-cost increase, and on the overall price index change (for all goods) compared to the price of the product. To me, this seems like a reasonable behavior equation.

I now complicate the analysis by considering the behavior equations that determine m_{1i} and m_{2i} when $F > \bar{F}$, and $P/p_i > (\bar{P}/p_i)$. Assume that m_{1i} and m_{2i} are "absorbtive" at low values of F and P, and further set

$$m_{1i} = f_1(F_i) \tag{4a}$$

$$m_{2i} = f_2(P/p_i) \tag{4b}$$

Let us assume that m_{1i} is first an increasing function of F_i, up to some maximum level, beyond which it is a decreasing function. Similarly, m_{2i} is a similar function of P/p_i. These ideas are illustrated in Figures 1 and 2. Note that, beyond the inert area bounds, m_{0i} is a constant, and since m_{1i} and m_{2i} have maximal values, then m_1 will have a maximal value. Further, since m_{1i} and m_{2i} decline beyond some point, note that m_i declines likewise. Aggregating for all sellers implies that M in the macro identity also declines beyond some maximal value.[5]

The case not covered is labor. Some who sell their labor will be weak in that market, while others, especially those in strong trade unions, will be strong. Since $F = 0$ for labor, the marginal markup equation for each worker j will be:

$$m_j = m_{0j} + m_{2j} \tag{4c}$$

To simplify matters assume that in the case of labor $m_{0j} = 0$. As before, to capture the inert area, assume that $m_{2j} = 0$ when $0 < P/p_j^w \leqslant \bar{P}/p_j^w$, and $m_{2j} > 0$ when $P/p_j^w > \bar{P}/p_j^w$ where p_j^w is the wage that j receives. Also,

$$m_{tj} = f_{tj}\left(\frac{P_t}{p_{tj}^w}, \frac{P_{t-1}}{p_{t-1,i}^w}\right) \tag{5}$$

[5] Since we are not interested in infinite inflation rates, let us assume that beyond some point $M < 1$. Space limitations do not permit, and it would take us too far afield, to establish the case for a maximum finite inflation rate. However, one could appeal to frictional elements and to a principle of increasing risk for loans.

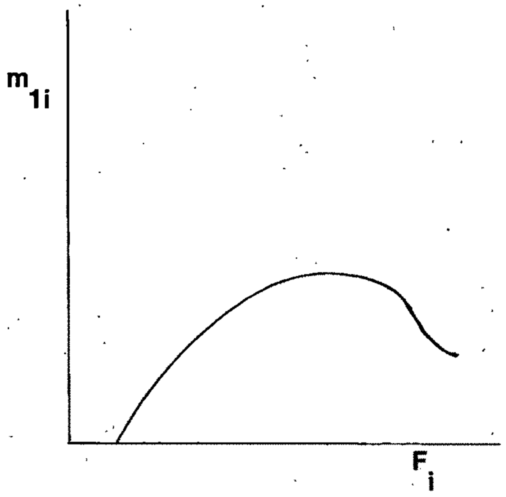


FIGURE 1

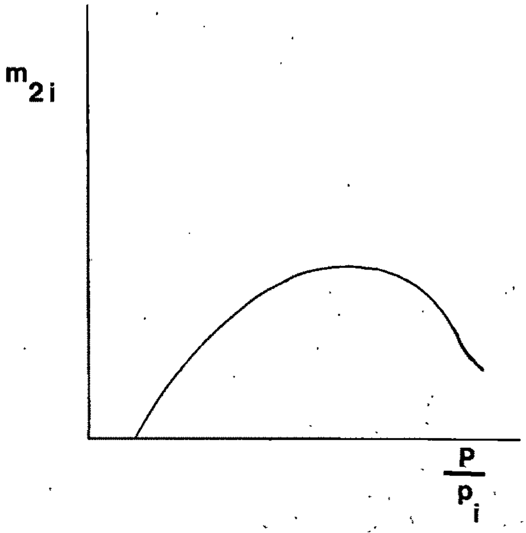


FIGURE 2

When labor is weak we would expect some labor to move to industries where it is strong. Therefore we take into consideration the previous period's ratio of the general price change to the wages change, assuming that this determines the extent to which labor moves. Equation (5) reflects whether labor is strong or weak in its market. In general if $P/p_j^w < 1$ labor is strong and if $P/p_j^w > 1$ labor is weak. Also m_{tj} will be high when labor is strong and low when it is weak.

Let us now turn to consider what happens when there is less than full accommodation of credit demanded. Consider the following relation:

$$A^* = A + S \tag{6a}$$

where A^* stands for the degree to which the increase in the money supply less than fully accommodates for the price increases, A stands for the degree to which less than complete accommodation reduces the multiplier M, and S is the extent to which a lack of accommodation reduces sales. In other words, the lack of accommodations results in part in a reduction of sales S and in part in a reduction of price. In the above it is possible that $A = .95$ while $(-S) = .05$. If sales fall then initially inventories build up, which in turn, after a lag, leads firms to reduce employment. In any event equation (1b) now changes to reflect less than full credit accommodation so that

$$M_t A_t P_t = P_{t+1} \tag{6b}$$

Of course A is an average which is distributed among a large number of firms in the following manner:

$$A = \sum_i a_i w_i \tag{6c}$$

where a_i is the adjustment on m_i because of a reduction in credit availability to the customers of firm i. Note that M and m_i represent the *ex ante* marginal price multipliers before they are adjusted for less than full credit accommodation.

Following equation (3a) we have a method of obtaining m_i for all i, from the relations (4a), (4b), and (4d) which determine m_{1i} and m_{2i}. Furthermore, by summing m_i according to equation (2c) we obtain values for M. Figure 3 indicates the macro relations that determine some of the equilibrium rates of inflation. The macro reaction to any value of P_t is $M_t A_t P_t = P_{t+1}$. The reaction curve is shown as R_1 in Figure 3. The 45° line marked E in the figure is the locus of equilibrium points under which $P_t = M_t A_t P_t = P_{t+1}$. Thus, the points of intersection be-

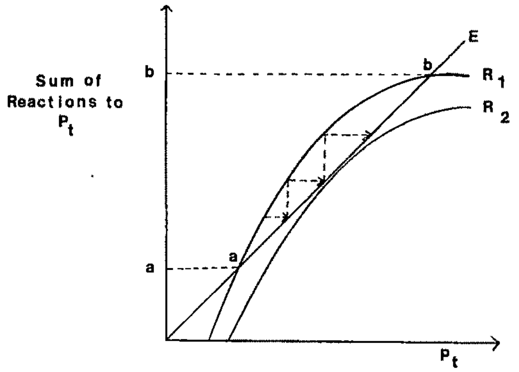


FIGURE 3

tween R_1 and the 45° line, i.e., a and b, are equilibrium values of inflation. However b is a stable equilibrium value and a is not. Starting with any inflation level above a will lead to a set of reactions that ends at b. Similarly for values above b we obtain a set reactions that lead to b. Values of P_t slightly lower than a generate a process that ends at zero inflation. Clearly, where the behavioral line is to the right of the 45° line reactions will be such as to decrease inflation and where R is to the left of E inflation will grow.

Figure 3 illustrates the effect of tightening the money supply by a rightward shift of the behavioral curve marked R_1 to R_2, and a loosening of the money supply by a shift from R_2 to R_1. If curve R_2 does not cut the 45° line we have a condition that is no longer consistent with sustained inflation. Any point on R_2 will generate a series of reactions that eventually leads to zero inflation. Hence my general approach can accommodate some monetarist ideas as well as administrative pricing notions.

V. Conclusions

I now turn briefly to some of the implications. 1) In accordance with the inert area principle some stimuli will be too small to generate sustained inflation. 2) By the same token some attempts to fight inflation already generated will be too small to be effective. The theory suggests that this possibility be considered in assessing current anti-inflation efforts. 3) Since some lack of credit accommodation results in part in sales reductions initially, attempts to fight inflation that have a direct influence on prices may cause less unemployment than those that work through indirect influences.

REFERENCES

J. W. Dean, "The Inflation Process: Why Conventional Theory Falters," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 362–67.

______**and H. Leibenstein**, *Towards a Micro-Behavioral Theory of Inflation*, forthcoming.

J. M. Fleming, *Inflation*, Oxford 1976.

Harvey Leibenstein, "A Branch of Economics is Missing: Micro-Micro Theory," *J. Econ. Lit.*, June 1979, *17*, 477–502.

______, *Beyond Economic Man*, Cambridge, Mass. 1976.

David Lewis, *Conventions*, Cambridge, Mass. 1969.

H. A. Simon, "Rationality as Process and as Product of Thought," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 1–16.

Increasing Unemployment and Changing Labor Market Expectations Among Black Male Teenagers

By LAURENCE C. MORSE*

For over twenty-five years there has been a secular divergence in the unemployment rates of black and white teenagers. Although the list of factors which may have contributed to the widening gap in the labor market experiences of black and white teenagers is quite long, there seems no reason, a priori, to believe that any one factor has operated in isolation, or with equal force, over the entire period. Suggested causes most frequently cited are: 1) differential growth rates in the black and white teenage populations; 2) an increasing difference in the employability characteristics of black and white teenagers, with black teenagers evidencing a decline in relative employability possibly due to widespread inadequacies in the quality of inner-city public schools; 3) employment decreasing effects of the minimum wage (although these should be race neutral in the absence of differences in employability or an employer preference for white over black youth); 4) an increase in labor market discrimination against black teenagers; 5) expectations among black teenagers with respect to wages and working conditions that have increased more than their attractiveness to employers; 6) increased competition for entry level jobs from other demographic groups (for example, adult white females); 7) the movement of industry out of central cities, reducing the number of entry level jobs available to residents left behind, for many of whom transportation to outlying areas is unavailable.

Of the many factors cited, two (increased discrimination and increased expectations) have been the subjects of comparatively little empirical work to assess their relative importance. This is due not so much, I believe, to an inherent lack of interest on the part of researchers, but rather to the fact that, on the one hand, discrimination in the labor market is believed to have lessened considerably, and, on the other, the increased expectations hypothesis appears not to be readily amenable to quantitative empirical examination. The focus here is on the increased expectations hypothesis.

In this paper I construct and present the results of estimating a model which I think gives us an "indirect" test of one dimension of the increased expectations hypothesis. Section I contains the conceptual discussion and model. Estimation results and some qualifications to these results are discussed in Section II.

I. An Indirect Test

A. *Discussion*

The basic notion underlying the increased expectations argument is as follows: over the past two decades, black youth cohorts have shown great increases in the quantity of education they have consumed. Increased education is thought to increase individuals' expectations and standards with respect to wages and job content which, when unmet by prospective employers, leads to refusal of employment offers.

Though some evidence in support of the hypothesis is to be found in the work of sociologists (see Elijah Anderson and Douglas Glasgow), this evidence falls short of a quantitative assessment of the relative

*Research associate, The Urban Institute. I would like to thank Orley Ashenfelter, Samuel L. Myers, Jr. and David H. Swinton for helpful comments on an earlier draft, and Joseph Altonji and Mark Plant for much valuable assistance. A longer version of this paper is available on request.

importance of increased expectations as a contributing factor to increasing unemployment among black teenagers. But how might one obtain such a quantitative assessment? We have here a sort of garden variety identification problem and in the absence of micro data on firms and individuals which would allow us to estimate "hedonic" wage regressions for two periods in order to determine the change in the implicit "psychic" costs to black youth of performing certain types of jobs, we seem to have no "direct" way of knowing quantitatively, just how important increased expectations (black pride, etc.) have been in causing black teenagers' unemployment rates to trend so sharply upward. Since data limitations prevent a direct test of the expectations hypothesis, the approach employed here is to verify the presence or absence of behavior, among black male teenagers, *implied* by the argument.

The increased expectations hypothesis *implies* that it is black teenagers with at least a high school education *who believe they have reason* to refuse some employment offers, having acquired what they have been told was the ticket to better jobs (here defined as jobs with pay and "status" comparable to those of their white counterparts with equal quantities of education) only to be met in the labor market by employers offering very low wages for tasks which provide little prestige, satisfaction or prospects for advancement. If the hypothesis is of much quantitative importance then, we should be able to observe a *lessening* of the difference between the effects of completing and not completing high school on the probability of being unemployed among successive cohorts of black male youth as successive cohorts of high school graduates increasingly refuse to accept jobs which are available to them. In Part B of this section, I construct a model of employment offer acceptance or refusal with which to test this particular implication of the increased expectations hypothesis.

B. *The Model*

First, I shall assume that each black male youth possesses a reservation wage which is a function of his years of schooling, amount of nonwork income, marital status, hours of labor supplied to the market, and the average market wage of white male youth with similar observable characteristics. More formally,

$$(1)\quad W_R^B = \alpha_0 + \alpha_1(ED) + \alpha_2(OI) + \alpha_3(MS) + \alpha_4(\overline{W}^W) + \alpha_5(H) + \varepsilon$$

where W_R^B is the reservation wage of a black male teenager, ED represents level of education, OI represents other income, MS represents marital status, $\overline{W}^W$ represents the average market wage of employed white male youth with similar observable characteristics, H represents hours of labor supplied to the market, and ε is an error term.

Second, each black male youth is assumed to face a market wage offer determined by the interaction of the supply of and demand for workers with similar observable characteristics. In the shortest possible period the labor market is assumed to have separate supply and demand curves at each level of schooling. More specifically, the demand for labor within schooling levels is a function of its price, the price of substitutes in production, the costs of other inputs in the production process (capital, raw materials) and the demand for the final product; the supply of labor within schooling levels is a function of the potential market wage, the comparative advantages of this type of labor in nonmarket activities, the average level of nonmarket income and the strength of the demand for labor. The labor market is assumed to clear and since the maximum supply of workers at any particular schooling level is fixed by the number of potential workers who have that amount of schooling, both market wages and employment/population ratios are determined for each schooling level. The reduced form equation for the market wage at a given schooling level may be expressed as

$$(2)\quad W_M^B = \beta_0 + \beta_1(ED) + \beta_2(\overline{W}^W) + \beta_3(\overline{OI}) + \beta_4(K) + \beta_5(Q) + \beta_6(U_{AM}) + \varepsilon$$

where K represents capital, Q represents demand for final product, U_{AM} represents the unemployment rate of adult prime-age male workers (here used as a proxy for the strength of the demand for labor), and all other variables retain designations previously assigned.

Third, a teenager's decision to accept or refuse an employment offer is assumed to be a direct function of the difference between the value of his individual reservation wage function at zero hours of labor supplied, $W_{R|h=0}$, and the market wage for individuals with similar observable characteristics, W_M. If $W_{R|h=0} < W_M$, the individual will accept the employment offer; if $W_{R|h=0} > W_M$, the individual will refuse the employment offer.

We have then two relationships, from which we may derive a third, that give us a testable model of the individual's decision to become employed or remain unemployed. Expressing the probability of an individual's being unemployed as a function of the difference between his reservation wage at zero hours of labor supplied and the market wage he faces, that is, the difference between equations (1) and (2), we may write

$$(3)\quad p(u)=\phi(W_R^B|h=0-W_M^B)$$

$$p(u)=\phi_0+\phi_1(ED)+\phi_2(OI)+\phi_3(MS)$$

$$+\phi_4(\overline{W}^W)+\eta$$

where $p(u)$ represents the probability of an individual's being unemployed.

Exactly how equation (3) allows us to perform an indirect test of the expectations hypothesis will be demonstrated with the aid of Figure 1, on the following page. In the shortest possible period the labor market is represented as having separate supply and demand curves at each level of schooling. The supply of workers at any particular schooling level is fixed by the number of workers who have that amount of schooling. Wages are represented on the vertical axis and people are arrayed along the horizontal axis according to years of schooling completed, lower schooling to the left. Demand curves are drawn indicating a willingness to hire increasing with years of school, but decreasing with the number of people. Supply curves slope up, as people offer their services with increasing wages, but become vertical at the limit of the number of people who have the indicated level of schooling.

Let us assume that Figure 1A represents a picture of the labor market for black male youths' 18 to 19-years of age, not enrolled in school in 1960. By allowing the demand curves for higher schooling categories to cross the supply curves at higher levels, I have explicitly assumed the validity of one of the basic predictions of human capital theory—that increased education reduces the probability of being unemployed. Let us further assume that Figure 1B represents a picture of the labor market for black male youth in 1970; here the proportion of workers with different schooling levels has been changed along the horizontal axis to show the greater proportion with twelve or more years of school. The 1960 labor supply and demand curves for black male youth with twelve or more years of school have also been redrawn in order to focus on the changes in their behavior and the behavior of employers over this period *implied* by the expectations hypothesis.

Focusing now only on the group with twelve or more years of school in Figure 1B, the supply curve has been shifted up, in keeping with the notion that at any given wage, a smaller proportion of black male youths were willing to accept employment in 1970 than in 1960. The demand curve for this group has also been shifted up but again, in keeping with the assumptions of the increased expectations hypothesis, this shift was not commensurate with the shift in the supply curve as employers' valuation of black male youth labor is assumed to have changed less than the youths' market valuation of themselves. Accordingly, the 1970 demand and supply curves intersect at point y at which wages are higher but unemployment is also higher due to the reduction in employment and the assumed extended search for acceptable jobs by young black males. What the expectations hypothesis *implies* is that the *difference* between the effects

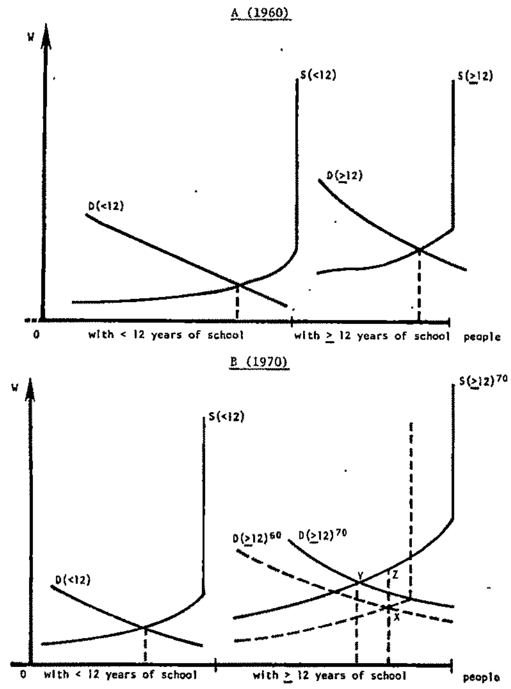

A (1960)
W
S(<12)
S(≥12)
D(≥12)
D(<12)
0
with < 12 years of school
with ≥ 12 years of school
people
B (1970)
W
S(≥12)70
S(<12)
D(≥12)60
D(≥12)70
D(<12)
Y
Z
X
0
with < 12 years of school
with ≥ 12 years of school
people

FIGURE 1

of having completed and having not completed twelve or more years of school on the probability of being unemployed should have been less in 1970 than in 1960 for black male teenagers; in terms of equation

$$(3)\quad p(u)=\phi_0+\phi_1(ED)+\phi_2(OI)+\ldots+\eta$$

where $ED=1$ if the individual completed twelve or more years of school and 0 otherwise, we should find that $(0-\phi_1^{70})<(0-\phi_1^{60})$ or simply $(-\phi_1^{70})<(-\phi_1^{60})$. In the specification of equation (3) used for the empirical work, ED becomes $ED1$ and/or $ED2$, representing less than twelve years of school and twelve or more years of school, respectively, and is interacted with a series of location dummies: S for South, W for West, and N for North/North Central regions. Equation (3) then becomes

$$\begin{aligned}p(u)=\phi_0&+\phi_1(ED1W)+\phi_2(ED1N)\\&+\phi_3(ED2S)+\phi_4(ED2W)\\&+\phi_5(ED2N)+\phi_6(OI)+\ldots+\eta\end{aligned}$$

When $(ED1S)$ is the excluded category, we should expect to find: $(0-\phi_3^{70})<(0-\phi_3^{60})$, $(\phi_1^{70}-\phi_4^{70})<(\phi_1^{60}-\phi_4^{60})$ and $(\phi_2^{70}-\phi_5^{70})<(\phi_2^{60}-\phi_5^{60})$.

C. *Data*

The data for the empirical investigation were extracted from the 1960/1970 Census Compatible Public Use Samples for black and white male youths who were 19 or 20 years of age not enrolled in school or serving in the armed forces in 1960 and 1970; as such, all earnings data and most employment data refer to the individuals' eighteenth or nineteenth year.

The primary dependent variable $U1$ is a dichotomous variable which takes the value 1 if the respondent did not work during the year prior to the survey and 0 if the respondent did work. Although this measure accords most accurately with the earnings and income data, covering the same time period, it is indisputably the case that many more individuals are unemployed at some time during a given year than are unemployed for the entire year and therefore a second variable, $U2$, unemployment status during the survey week, was also used as a dependent variable; $U2$ is also a dichotomous variable assigned the value 1 if the respondent was unemployed during the survey week, 0 otherwise.

II

A. *Results and Conclusions*

Maximum likelihood estimates of the coefficients in equation (3), with $U1$ as dependent variable, were obtained using a standard probit estimation package. Changes in the difference between the effects of completing and not completing high school on the probability of being unemployed over the period 1960 to 1970 were then computed for black and white male youths. For black male youths, in not a single instance was the difference in the effects of education level on the probability of being unemployed less in 1970 than in

1960. On the contrary, the difference by region was consistently greater in 1970 than in 1960 although never significantly so. The changes in the differences for the South, West, and North/North Central regions were .186 (1.370 *t*-statistic), .220 (.489), and .337 (1.617), respectively. For white male youths, the exact opposite pattern of results prevailed, that is, the difference in the effects of education level on the probability of being unemployed were consistently *less* in 1970 than in 1960 by region, a pattern of results consistent with behavior implied by the expectations hypothesis. Again, however, the changes between 1960 and 1970 by region were never significant at the 5 percent level, though the change for the southern region was significant at the 10 percent level. The changes in the difference for the South, West, and North/North Central regions were −.154 (−1.793), −.082 (−.521), −.138 (−1.225), respectively. Results obtained using *U2* as dependent variable did not alter the pattern or significance of results for either race group.

In summary, the results failed to confirm a significant decrease in the difference in the effects of completing and not completing high school on the probability of being unemployed for either race group. Although insignificant in the statistical sense, the results displayed consistent *decreases* by region between 1960 and 1970 for white male youths and consistent *increases* for black male youths. The findings give no support to the presence of a quantitatively significant change in the behavior of black male youths consistent with the increased expectations hypothesis.

B. *Qualifications*

The results presented above should be weighted by the following qualifications, all of which would tend to bias the results against finding significant change among black male youth. First, the dependent variable does not distinguish between (a) individuals who seek jobs and are refused, (b) individuals who seek jobs and refuse those offered to them, and (c) individuals who—believing that they will not be offered the jobs they want or will not accept the ones they are offered—do not seek jobs. The unemployment probability measure includes (a) and (b) but excludes (c). Although I have assumed market clearing via equilibrium wages, several researchers have noted that young blacks tend to leave the labor force more often than young whites when they are without jobs and are therefore more often not counted among the unemployed. Second, to the extent that the behavioral change that we are seeking to verify is the result of the interaction of the Civil Rights Movement (on black pride) and the stay in school campaign of the 1960's, the behavioral changes might be more evident between cohorts entering the labor market in the mid- to late-1960's and those entering the labor market in the mid- to late-1970's. The use of 1960/1970 data may be a severe handicap. Analysis of 1970/1980 data may reveal significantly different results. Third, since the connection between level of educational attainment and an individual's actual ability to perform a given task is not clear, it is also not clear that the supposed behavior alteration among black male youth should be thought to have occurred only among high school graduates; the attitudes of black male high school dropouts might also have changed. Finally, the work presented here has focused on only one dimension of the prospective employee's job acceptance decision—wages. To the extent that such factors as opportunity for advancement and prestige are equally important elements of black youths' job expectations, this singular approach may miss the most important dimensions along which black youths' expectations have altered. I hope to address these considerations in future work.

REFERENCES

Elijah Anderson, "Some Observations on Black Youth Employment," in Bernard E. Anderson and Isabel V. Sawhill, eds., *Youth Employment and Public Policy*, Englewood Cliffs: Prentice-Hall, Inc. 1980.

Douglas Glasgow, *The Black Underclass: Poverty, Unemployment and Entrapment of Ghetto Youth*, San Francisco: Jossey-Bass Publishers 1980.

Federal Minimum Wage Laws and the Employment of Minority Youth

By CHARLES L. BETSEY AND BRUCE H. DUNSON*

Numerous studies have emerged over the past decade dealing with the employment effects of minimum wage legislation (for example, see Jacob Mincer and Edward Gramlich). These studies uniformly show that some amount of disemployment results from the imposition of a minimum wage above a wage that would prevail in the absence of such legislation; though there is considerable dispute concerning the magnitude of the effect.

In addition, most if not all of the recent studies indicate a greater employment sensitivity to changes in the minimum wage on the part of teenagers, nonwhites, and the low-skilled, generally, than is true for other groups. Job losses due to higher minimum wages may contribute to high unemployment rates among youth in general, and minority youth in particular.

This paper presents no direct evidence on the impact of the minimum wage on teenage unemployment. Its results do suggest, however, that previous studies of the impact of the minimum wage on employment loss were probably biased towards overstatement of the effect of the minimum wage. In the case of nonwhite teenagers, the results presented here indicate that over the period 1954 to 1979, cyclical factors affected nonwhite teenage employment to a far greater extent than did changes in the minimum wage.

In fact, over this same period there is no discernible disemployment effect associated with higher minimum wages for nonwhite teenagers, *ceteris paribus*. While during the most recent period (1970–79), there is clear indication that changes in the minimum wage resulted in reduced employment for nonwhite 16–19-year olds.

In Section I, we discuss the bias in previous studies resulting from the omission of a control for income transfer programs. The results of two earlier studies that attempted to control for these effects are discussed in Section II. Section III proceeds with a discussion of our empirical results, followed by a summary.

I. Income Transfer Programs and Bias

The issue of whether or not an income effect is important in determining labor supply response, depends, among other things, on its size. The results of the various income maintenance experiments, as well as those of statistical studies, indicate that there are significant income effects associated with income transfers that result in reductions in work effort. The reductions are particularly large for wives and, according to a recent study by Richard West, young nonheads of household. According to West, income supplements for young male nonheads in the experiment resulted in reducing average hours worked per week by 4.6 hours or about 24 percent, and the proportion of time spent working was reduced by about 21 percent. The reductions for females were somewhat less strong.

The finding of a negative income effect for young nonheads makes ignoring these effects in minimum wage studies particularly troublesome. Given the low wages generally earned by youths, and the prevalence of income transfer programs for which many may be eligible (either as direct recipients or

*The National Urban Coalition, and the University of Maryland, respectively. Funding for this research was provided in part by a grant from the Ford Foundation to the National Urban Coalition. We wish to thank Charles Brown and Curtis Gilroy of the Minimum Wage Study Commission for providing access to the data used in the study. Comments were provided by Bernard Anderson, Bruce Vavrichek, Alan Fechter, George Iden, Daniel Saks, Fred Siskind, and David Swinton. The opinions expressed here are solely our own and do not necessarily reflect the opinions of the organizations with which we are affiliated, or of their funding agencies.

as part of a larger family unit), ignoring the income effects associated with income transfer programs may significantly bias previous estimates of the impact of minimum wage increases on youth disemployment.

In the simplest possible case, labor force status is a function of a minimum wage (x_1) and other income (x_2). A reduced-form equation is estimated that, however, excludes income. The true equation is

$$Y = b_1 x_1 + b_2 x_2 + u \tag{1}$$

The estimated equation, however, is

$$Y = b_1{}^* x_1 + u^* \tag{2}$$

where

$$\hat{b}_1{}^* = \frac{\Sigma x_1 y}{\Sigma x_1^2} \tag{3}$$

or

$$\hat{b}_1{}^* = \frac{\Sigma x_1 (b_1 x_1 + bx + u)}{\Sigma x_1^2} \tag{4}$$

Thus

$$E(\hat{b}_1{}^*) = b_1 + b_2 \frac{\Sigma x_1 x_2}{\Sigma x_1^2} + \frac{\Sigma x_1 u}{\Sigma x_1^2} \tag{5}$$

Since the expected value of the last term is zero, we have $E(\hat{b}_1{}^*) = b_1 + b_2 b_{12}$ where b_{12} is the regression coefficient in the "auxiliary" regression of the excluded variable X_2 on the included variable X_1. The bias observed equals the true coefficient of the omitted variable times the regression coefficient of the excluded variable on the included variable.

In empirically examining the magnitude of the bias that might result from the omission of a control variable for income effects, it is assumed that changes in both the minimum wage variable and the benefit level of transfer payments are part of a larger system of equations that affect labor supply.

II. The Minimum Wage and Transfer Payments

Several previous studies have attempted to examine the relationship between youth employment and the welfare system. Of these studies, those by T. Kelley, and Michael Wachter and C. Kim are the closest to what is attempted in this paper.

In a series of two papers, Kelly develops a model to evaluate the hypothesis that a minimum wage adversely affects the employment of teenagers. A residualization technique is employed in estimating his model. For the female equations Kelly makes use of this residualized welfare variable. It was found to have no appreciable effect on the estimated impact of the minimum wage. In constructing this welfare variable, Kelly used the total number of recipients in the Aid to Families with Dependent Children (AFDC) program. As reported by C. Brown et al., the residualization of this variable effectively guarantees that the estimated minimum wage impact will not be appreciably affected by including this welfare variable.

In their study, Wachter and Kim argue that taking account of the impact of government social welfare programs is essential to the study of the labor force behavior of youth. They point out that youth tend to have a relatively low attachment to a given employer or the labor force. Furthermore, teenagers are often secondary wage earners. As a consequence, these workers are closer to the margin of working and not working. The result is that increases in the levels and coverage of transfer payments would be expected to increase the duration and frequency of their unemployment spells. As an empirical proposition, Wachter and Kim point out that changes in minimum wage levels and coverage have to some extent coincided with changes in public assistance programs. If, for example, as programs such as AFDC expand, we also observe lower employment and labor force participation among youth; previous studies may have incorrectly attributed these effects to changes in the minimum wage. Wachter and Kim's own attempts to control for changes in welfare programs produced insignificant results.

III. Results

Since our concern is with the possibility of a bias resulting from the omission of a welfare variable, it was decided to duplicate

TABLE 1—MEANS AND STANDARD DEVIATIONS OF SELECTED VARIABLES[a]

Variables	Means	Standard Deviations
White teens (16–19), E/P	48.90	7.20
Nonwhite teens (16–19), E/P	32.94	5.54
Minimum Wage, Teens	25.28	7.05
Armed Forces, Teens	4.52	1.06
Cyclical variable, UC	3.28	1.31

[a] Shown in percent.

a study that was fairly simple, yet relatively complete. The model chosen was originally presented by Masanori Hashimoto and Mincer, and in Mincer. Although their sample was disaggregated into ten separate age, race, and sex strata, we only consider the white and nonwhite teen category. Our analysis is further limited to consideration of the employment to population ratio of the ith demographic group in time period t as the dependent variable.

The empirical model reported by Mincer was of the form $Y = F$ (MW, AF, UN, T, T^2). The minimum wage variable (MW) is the familiar Kaitz index. It is the ratio of the minimum wage to average hourly earnings multiplied by industry teenage employment and the proportion of total employment covered under the Fair Labor Standards Act. Since firms might be expected to adjust to past increases with a lag, the minimum wage variable is entered into the regressions in a distributed lag form. Specifically, an Almon unconstrained quadratic distributed lag, with a lag pattern of eight quarters is used.

The adult male unemployment rate (UN) age 45–54 expressed quarterly is used as a proxy for the business cycle. This variable is considered by many to be a reliable measure, in large part because the labor force participation rate of this group is relatively cyclically insensitive. The Armed Forces variable, AF, is the ratio of the total male population 16–19 in the armed forces to total teen population. The final variable in Mincer's empirical model was a time trend, which is a crude substitute for a more complete specification of employment and labor force functions.

Based on the model presented in Hashimoto and Mincer, we obtained results for the minimum wage variable that are not significantly different from those presented in Mincer. The minor differences that remain can be explained by the fact that our minimum wage variable was based upon actual quarterly data, while the Hashimoto-Mincer quarterly index was generated by interpolating annual data, taking into account the trend in the minimum wage variable and the timing of changes in nominal wages. We both use an employment variable unadjusted for seasonal variation.

The means and standard deviations for our variables are presented in Table 1. Table 2 presents our estimated results compared to the Hashimoto-Mincer results for the period 1954I to 1969IV. As can be seen, the results for white teens are virtually identical while the results for nonwhite teens exhibit some differences.

The equations presented in Table 3 include a measure of welfare transfer payments. The measure used is the average per person payment for AFDC. This variable was deflated by the implicit GNP price deflator to constant 1972 dollars. The welfare variable is highly significant for white teens. Moreover, its effect on the minimum wage coefficient is as hypothesized. That is, the sum of the lagged minimum wage coefficients remain significantly different from zero.

For nonwhite teens we also observe a similar effect. The magnitude of the minimum wage coefficient drops, but remains statistically significant. Unlike the results for white teens, however, the welfare variable for nonwhite teens is not significantly different from zero.

Since data were available up to the fourth quarter of 1979, we decided to update our estimated equations. Initially we did this by estimating the equations presented in Table 3 for the time period 1970–79. The results are presented in Table 4. For white teens, a number of changes are observed. First, the sensitivity of this group to cyclical changes in the economy becomes rather obvious. The unemployment coefficient increased to - 1.81. The disemployment effect of the minimum wage did not change significantly.

TABLE 2—DISEMPLOYMENT EFFECTS OF MINIMUM WAGES ON WHITE AND NONWHITE TEENS

	Hashimoto-Mincer Results		Replication of Hashimoto-Mincer	
	White	Nonwhite	White	Nonwhite
$t-0$	$-.078$	$-.009$	$-.064$	$.036$
$t-1$	$-.051$	$-.027$	$-.045$	$.022$
$t-2$	$-.033$	$-.046$	$-.033$	$.002$
$t-3$	$-.022$	$-.065$	$-.027$	$-.024$
$t-4$	$-.020$	$-.084$	$-.029$	$-.057$
$t-5$	$-.027$	$-.102$	$-.038$	$-.095$
$t-6$	$-.041$	$-.121$	$-.053$	$-.140$
$t-7$	$-.064$	$-.140$	$-.076$	$-.193$
$Sum\,(Mw)_L$	$-.336$	$-.594$	$-.367$	$-.447$

TABLE 3—THE DISEMPLOYMENT EFFECTS OF THE MINIMUM WAGE, CONTROLLING FOR WELFARE PROGRAM EFFECTS

	White Teens		Nonwhite Teens	
Variables[a]	Coefficients	t-statistics	Coefficients	t-statistics
Constant	.7666	9.46	.595	3.36
T	$-.005$	-6.91	$-.008$	-4.68
T^2	.0001	8.58	.0001	3.93
AF	$-.038$	$-.297$	$-.242$	$-.843$
$D62$	$-.017$	-2.19	.012	.700
$D2$	.056	8.36	.058	3.87
$D3$	.125	23.8	.111	9.61
$D4$	.031	6.77	.045	4.55
GS	$-.439$	-2.76	$-.103$	$-.297$
$Sum(UN)_L$	$-.924$	-3.57	$-.482$	$-.104$
$Sum(MW)_L$	$-.259$	-2.73	$-.420$	-2.00
R^2	.98		.90	
DW	1.66		1.74	

[a] All the variables are as previously defined with the exception of GS which is the average per person payment for AFDC.

TABLE 4—THE DISEMPLOYMENT EFFECTS OF THE MINIMUM WAGE OVER DIFFERENT TIME PERIODS

	White Teens				Nonwhite Teens			
	1970–79		1954–79		1970–79		1954–79	
Variables[a]	Coefficients	t-statistics	Coefficients	t-statistics	Coefficients	t-statistics	Coefficients	t-statistics
Constant	.363	7.23	.377	8.86	3.06	3.64	.451	6.54
T	$-.003$	-7.27	$-.002$	-2.64	$-.061$	-2.99	$-.003$	-3.30
T^2	.00001	15.35	.0003	8.96	.0003	3.03	.00002	3.46
AF	.603	.846	.333	2.30	1.25	.703	$-.054$	$-.234$
$D62^1$			$-.034$	-4.40		1.97	$-.015$	-1.20
$D2$	.054	9.09	.063	9.51	.026	8.59	.039	3.67
$D3$	.093	21.2	.114	22.4	.095	1.71	.091	11.1
$D4$	.016	5.43	.022	5.09	.013	1.31	.021	3.08
GS	.065	3.23	.168	1.90	.276	.105	$-.006$	$-.045$
$Sum(UN)_L$	-1.81	4.80	-1.99	12.28	$-.073$	3.14	-1.98	-7.52
$Sum(MW)_L$	$-.222$	.770	$-.150$	2.12	-1.66		$-.033$	$-.289$
R^2	.96		.96		.96		.91	
DW	.94		1.17		1.57		.159	

[a] The variable $D62$ was dropped over the time period 1970–79 since it was one over this entire period.

At −.222, however, its effect was not statistically significant from zero. The sign on our welfare measure was positive and not significantly different from zero.

The period 1970−79 exhibits major changes in the labor market for nonwhite teens. In fact, the major results for nonwhite teens are just the opposite of those observed for white teens. Over this time period, nonwhite teens do not appear to be affected by cyclical movements in the economy, as indicated here by the low t-statistic on our unemployment variable. The disemployment effect of the minimum wage, however, increases dramatically for nonwhite teens.

This result contradicts earlier studies by Marvin Kosters and Finis Welch that indicate a greater cyclical sensitivity for nonwhites based on discrimination and lower skill levels. On the other hand, the result is consistent with George Iden's finding that the E/P ratio for blacks tends to be less cyclically sensitive during periods he classifies as "high unemployment" periods than during periods classified as "low unemployment" periods. The 1970's were a high unemployment period.

We next estimated the same equations over the time period 1954–79. The results are in many respects consistent with those of other investigators. Differences however, did exist: both the armed forces variable and our welfare variable had positive rather than negative coefficients in the case of white teens and were significantly different from zero. The total effect of unemployment was −1.99 and significant. The minimum wage coefficient was −.150 with a t ratio of 2.14.

The results for nonwhite teens, when compared with white teens, are consistent with previous findings. To begin with, the AF variable is of the expected sign. Similarly our newly introduced welfare variable has a negative coefficient, though its effect is not significantly different from zero. The major finding for the 1954–79 period, that in the case of nonwhite teens cyclical factors are more important than increases in the minimum wage for determining employment changes, is consistent with earlier studies.

Our results for the recent 1970–79 subperiod indicate a statistically significant large disemployment effect for nonwhite teens. While there was also a significant disemployment effect associated with the minimum wage for white teenagers during the 1970's their employment was not nearly as sensitive to the minimum and considerably more sensitive to cyclical factors than nonwhite teenagers' employment.

The results presented here indicate striking differences in the disemployment effects associated with changes in the level and coverage of the minimum wage for nonwhite teenagers compared to white teens in the 1970s. Our results also indicate that changes in the overall state of the economy had less of an impact on black teens' employment prospects during the high unemployment era of the 1970's than in previous time periods. In addition, disregarding the effect of welfare programs has led to biased estimates of the effect of minimum wage increases in earlier studies.

Further refinement of the welfare variable is required, as is a more systematic investigation of the sectoral impacts of increases in the minimum wage. The measured effect of changes in the minimum contains errors to the extent that shifts from full-time to part-time work or other adjustments in hours of work occur in response to a change in the minimum. For that reason, a measure of hours worked would be a more appropriate dependent variable than the standard employment-population ratio.

REFERENCES

Bernard Anderson and Isabel Sawhill, *Youth Employment and Public Policy,* Englewood Cliffs: Prentice-Hall 1980.

C. Brown et al., "Effects of the Minimum Wage on Youth Employment and Unemployment," work. paper no. 1, Minimum Wage Study Commission, May 1980.

E. Gramlich, "Impact of Minimum Wages on Other Wages, Employment and Family Incomes," *Brookings Papers,* Washington 1976, *2*. 409–61.

M. Hashimoto and J. Mincer, "Employment and Unemployment Effects of Minimum Wages," unpublished paper, Nat. Bur. Econ. Res., 1970.

G. Iden, "The Labor Force Experience of

Black Youth: A Review," *Monthly Labor Rev.*, Aug. 1980, 10–16.

T. Kelley, "Youth Employment Opportunities and the Minimum Wage: An Econometric Model of Occupational Choice," unpublished paper, The Urban Institute, 1975.

M. Kosters and Finis Welch, "The Effects of Minimum Wages on the Distribution of Changes in Aggregate Employment," *Amer. Econ. Rev.*, June 1972, *62*, 323–32.

Stanley Masters and Irwin Garfinkel, *Estimating the Labor Supply Effects of Income Maintenance Alternatives*, New York: Academic Press 1977.

J. Mincer, "Unemployment Effects of Minimum Wages," *J. Polit. Econ.*, Aug. 1976, *84*, S87–S104.

D. Saks, *Public Assistance for Mothers in An Urban Labor Market*, work. paper no. 118, Industrial Relations Sec., Princeton Univ. 1975.

F. Siskind, "Minimum Wage Legislation in the United States: Comment," *Econ. Inquiry*, Jan. 1977, *15*, 135–38.

M. Wachter and C. Kim, "Time Series Changes in Youth Joblessness," in R. Freeman and D. Wise, eds., *Youth Unemployment: Its Nature, Causes and Consequences*, forthcoming.

F. Welch, "Minimum Wage Legislation in the United States," *Econ. Inquiry*, Sept. 1974, *12*, 285–318.

______, "Minimum Wage Legislation in the United States: Reply," *Econ. Inquiry*, Jan. 1977, *15*, 139–42.

Richard West, *The Effects of the Seattle and Denver Income Maintenance Experiments on the Labor Supply of Young Nonheads*, Menlo Park: SRI International, May 1979.

FREIGHT TRANSPORT REGULATION AND MARKET IMPERFECTIONS

Market Structure and Concentration in the Regulated Trucking Industry

By RUSSELL C. CHERRY AND CARL BACKMAN*

This paper presents estimates and interpretations of the Industry Performance Gradient Indexes (*IPGI*) developed by Robert Dansby and Robert Willig (hereafter, D-W) using estimated data for the regulated, general freight motor common carrier industry, hereafter "the trucking industry." These indexes measure the social welfare performance of a given industry structure and represent the difference between price and marginal cost as a proportion of price. In a related measure, they also give the welfare gain of governmental intervention in the industry. Both types of indexes are shown to be transformations of market share and price elasticity. For further discussion consult D-W. The formulas for the indexes are shown in the Appendix; the indexes themselves will be discussed, but the formulas are not shown in the body of the text.

Trucking firms set and publish rates collectively through organizations called motor rate bureaus. The bureaus have antitrust immunity under the Reed-Bulwinkle Act of 1948, with the nominal oversight of the Interstate Commerce Commission (ICC). In practice, the ICC exercises little control over ratemaking. The ICC also controls entry through the grant of operating authority, or "certificates," that list specific routes that carriers may serve. Grants of authority with extensive geographical coverage have traditionally been virtually impossible to obtain. They were usually assembled by a process called "tacking" or assembling individually purchased authority piece by piece.

I. Data and Assumptions

The data that were used to compute the *IPGI* are a freight bill sample that is stratified by carrier and weight-bracket, collected annually, and processed by each of the motor rate bureaus. The underlying sample design is statistically sophisticated and well thought out, and provides precise and unbiased estimates of revenues, shipments, and tons carried for participating carriers. This sample is called the Continuing Traffic Study (*CTS*) and it usually represents approximately 85 percent of general freight revenues.

A subset of the *CTS* was used to compute the origin and destination (*OD*) market shares and price elasticity of demand estimate needed to compute the *IPGI*. This subset consists of *OD* pairs with 50,000 or more expanded-sample shipments in 1976–77; there were 442 such *OD* pairs. The assumption underlying this subset is that the relevant market is a single *OD* pair.[1] The *IPGI* were then computed using this *OD*

*Arthur D. Little, Inc. and Senate Computer Center, respectively. We acknowledge the support of the Senate Judiciary Committee, Senator Edward M. Kennedy, Chairman, and the Committee Staff, especially Jay Steptoe, Jeff Stone, David Boies, Stephen Breyer, and Jim Michie. We acknowledge useful communications from W. Edwards Deming and Benjamin Tepping on statistical bias in the estimation of concentration. We are grateful for comments from Ann Friedlaender, Annette Lamond, and to Bob Dansby for pointing out an error in an earlier version of the paper. We acknowledge the programming assistance of Jim Estep, Craig Boyle, and Kathy Gust. Any remaining errors are our sole responsibility.

[1] W. Edwards Deming and Benjamin Tepping in private communications have pointed out an upward bias in estimates of concentration from sample data. This problem arises for a problem in the ranking and selection of firms by size. A mathematical expression for this bias is not available, because it depends on both the true unknown rank of the firm as well as the estimated rank. Simulations show that the average bias for the corridors in the *LTL* sample are 5 percent under plausible assumptions about the true size distribution of firms.

pair data. The price elasticity of demand was estimated with a *log*-linear regression of price on quantity. See Cherry for more detail on the contents of the *CTS*.

The *IPGI* allow us to infer price marginal cost relationships without observing actual marginal cost, with one exception. The *IPGI* assume a homogeneous product and this is not uniformly appropriate for trucking because of dichotomy in the type of service offered by carriers: if the shipment tendered weighs, say, 10,000–15,000 pounds, the shipment is classified as "truckload" (*TL*), if less, it is a "less-than-truckload" (*LTL*) shipment.

Carriers that specialize in *LTL* perform a more involved service for shippers than *TL* carriers. The *LTL* consists of: 1) initial shipment pick up; 2) terminal activity (shipment aggregation and sorting by destination); 3) line-haul activity; and 4) delivery at the final destination. Shipments may also undergo an intermediate disaggregation called a "break bulk." Truckload movements consist of only the line-haul service described above. Only freight bills for *LTL* shipments were used to compute the indexes so as to satisfy the homogeneous product assumption.[2]

Since trucking demand is derived demand, we must make one additional assumption. We assume that the sectors of the economy between shippers and final consumers are perfectly competitive so that the benefits to reallocating current output through marginal cost pricing would pass through to the final consumer.

II. The *IPGI* and *VTI*

Dansby and Willig derive five indexes and five corresponding values to intervention (*VTI*). Each *IPGI* has a corresponding *VTI* with the same behavioral assumption. All of the indexes are derived from the equilibrium condition for an oligopolistic firm. The *IPGI* are denoted by Φ_i, $i=1-5$; the subscript denotes the underlying behavioral assumption. The *VTI* are denoted by $\Phi_i^{\$}$.

The indexes are: 1) Φ_1, assumes that the firms are profit-maximizing oligopolies with some estimate, a conjectural variation, γ_i. The conjectural variation is estimated following Gyoichi Iwata; 2) Φ_2, assumes that oligopoly firms are quantity-Cournot with zero conjectural variations; 3) Φ_3, assumes that there are m large quantity-Cournot oligopolists with a competitive fringe of smaller firms that are price takers; 4) Φ_4, assumes that the largest m firms jointly maximize profit (collusively) with the remaining firms constituting a competitive fringe.

The derivation of the index itself involves solving the equilibrium condition for price less marginal cost divided by price as a function of price elasticity and market shares, such as the Herfindahl Index, a truncated Herfindahl, or the m firm concentration ratio. For example, the index Φ_1 is a function of the Herfindahl Index times one plus the conjectural variation squared, while Φ_4, assumes that firms are collusive and is a function of the m firm concentration ratio.

A fifth index assumes that there are differentiated products for each firm and an elasticity for each product. This index was not computed because we do not accept the assumption that *LTL* output is differentiated by firm.

Note that, strictly speaking, indexes Φ_1, Φ_2, Φ_3, and Φ_5 represent behavioral assumptions that are not satisfied for the trucking industry. Since rates are set through the motor rate bureaus and they may not legally be varied without going through the rate bureau process, the implication of variable prices and quantity associated with oligopoly interaction are not strictly applicable. However, the nominal rate that we use as price in this exercise is not the only component of the shipper costs: The "full price" (see Gary Becker) of the freight movement is the nominal rate plus the value of time to the shipper. This full price is a relevant benchmark for the indexes discussed here; since it can be varied by firm, through

[2] There is a distinction that should be made between "truck full" and "truck load." Truck load is based on a distinction that is made usually between 10,000-15,000 pounds in a tariff. On the other hand, a full truck of *LTL* freight averages around 28,000-33,000 pounds.

TABLE 1—DESCRIPTIVE STATISTICS FOR INDUSTRY PERFORMANCE INDEXES

	Mean	Standard Deviation	Sum of Values to Intervention	Value of Sum/ Total Revenues
Φ_2	.2990	.0784		
Φ_3	.2797	.0899		
$\Phi_2^\$$	334,896	360,434	148,011,877	.0688
$\Phi_3^\$$	\$1,170,338	922,371	517,289,428	.2406

service competition, firms may interact as oligopolists even though the nominal rate is constant. The service competition phenomenon has long been recognized as an important one in trucking. Also, if one includes a quality variable in the demand function, then the index for a quantity-Cournot oligopolist is invariant. That is, the index Φ_2 can be shown to be the same with the inclusion of quality as without. (See Dansby p. 150.)

III. Estimates of *IPGI* and the Value to Intervention

While we do not report them here, we did estimate the *IPGI* associated with Φ_1 that assumes a conjectural variation. This requires an estimate of marginal cost, that was obtained by assuming that marginal cost is equal to average cost. As a sensitivity test, we vary marginal cost up and down and the computed the conjectural variation and the associated *IPGI*, Φ_1.

We do not believe that this is the correct index for trucking because it implies a degree of sophistication about competitors' behavior that is alien to the trucking industry. The industry is far from having the degree of sophistication that one might associate with many oligopolies such as the automobile industry, the oil industry, or the breakfast cereal industry.

The two best candidates for the trucking *IPGI* are Φ_2 and Φ_3: Φ_2 assumes that firms are quantity-Cournot, while Φ_3 assumes that there are *m* large quantity-Cournot price leaders with a price-taking fringe. Descriptive statistics on these indexes are displayed in Table 1. The mean value for Φ_2 is .2990, while the mean value for Φ_3 is .2797. The table also displays values for the standard deviation, coefficient of variation, skewness, and kurtosis of the two *IPGI*. It is difficult to choose which index is most appropriate because the correct one probably varies by corridor. But because we must choose one index, Φ_3 sums best principally because what occurs in an individual market is that the major carriers in that market, which are not necessarily the major industry carriers, set rates and service options while the remaining carriers follow suit. Note that Φ_3 indicates a substantial gap between price and marginal cost without an estimate of marginal cost, and hence is more likely to be correct then Φ_1, which does require marginal cost. In addition, Φ_1 requires a behavioral assumption that we argue incorrect.

IV. Conclusions

The index that represents the most likely behavioral assumption is Φ_3 which had a mean value of .2797. This implies that the across-firm average, proportional difference between price and marginal cost is 28 percent of the average rate, for *LTL* general freight carriers in major *OD* markets.

What does this tell us about the industry as a whole? We believe that Φ_3 would be smaller for *TL* firms because there is some empirical evidence of a lower markup in *TL*. In any case, we cannot apply the value of any index to industry revenue and estimate the welfare gain to marginal cost pricing. (See D-W's Proposition 4, p. 253.) In order to find the upper limit to welfare gain we must apply some metric (denoted by ρ in D-W) to adjust this index to account for the movement toward optimal output, with marginal cost pricing.

The *VTI* must also be adjusted to account for the change in output that would result when prices were moved toward, or equal to, marginal cost. This procedure involves using a metric to adjust the *VTI*

$$(1)\qquad \rho(\Delta q_1 q^0)=\left[\Sigma(\Delta q_i/q_i^0)^2\right]^{1/2}$$

The calculation of the metric requires an estimate of the square root of the percentage change in quantity squared for each carrier. This metric has not yet been computed for the value to intervention. As a consequence the value to intervention presented does not yet represent the upper limit of the welfare gain to government intervention.

V. Summary

The *IPGI* allow us to infer proportional price marginal cost differences. We have shown that these have a mean value of .2797 using the index that best suits the trucking industry, Φ_3. No caveat need be associated with the use of this index.

On the other hand, the *VTI* require adjustment by a metric ρ, which was not calculated in the initial estimate of the *VTI*. As a consequence, the values of $\Phi_3^{\$}$ presented would have to be adjusted by the metric before they properly represent the upper limit to the value of intervention.

The computation of indexes for firms with a quality variable in the demand function, deserves some additional attention. In particular the quality dimension, in general, could be expected to increase the value of elasticity and reduce the appropriate *IPGI*.

Appendix: Formulas for Industry Performance Gradient Indexes and Values to Intervention

IPGI

$$\Phi_1=1/\xi\left[\sum_{i=1}^{n} S_i^2(1+\gamma_i)^2\right]^{1/2}$$

$$\Phi_2=1/\xi\left[\sum_{i=1}^{n} S_i^2\right]^{1/2}$$

$$\Phi_3=1/\xi\left[\sum_{i=1}^{m} S_i^2\right]^{1/2}$$

$$\Phi_4=\sqrt{m}\,/\xi\left[\sum_{i=1}^{m} S_i\right]$$

$$\Phi_5=\left[\sum_{i=1}^{n}(1/\xi_i)^2\right]^{1/2}$$

ξ_i = Elasticity for *i*th product

VTI

$$\Phi_1^{\$}=PQ/\xi\left[S_i^4(1+\gamma_i)^2\right]^{1/2}$$

$$\Phi_2^{\$}=PQ/\xi\left[\sum_{i=1}^{n} S_i^4\right]^{1/2}$$

$$\Phi_3^{\$}=PQ/\xi\left[\sum_{i=1}^{m} S_i^4\right]^{1/2}$$

$$\Phi_4^{\$}=PQ/\xi\left[\sum_{i=1}^{n} S_i\right]\left[\sum_{j=1}^{m} S_i^2\right]^{1/2}$$

$$\Phi_5^{\$}=PQ\left[\sum_{i=1}^{n}(R_i/\xi_i)^2\right]^{1/2}$$

$R_i=P_iQ_i/\sum_{i=1}^{n}P_iQ_i$ Revenue share of *i*th firm

where S_i = output share of *i*th firm in ton miles; γ_i = conjectural variation $\gamma_i=\Sigma Q_i/Q_i$ $i\neq j$; p = rate (price); Q = output in ton miles; and ξ = elasticity of demand.

REFERENCES

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

R. C. Cherry, *An Analysis of the 1976 Continuous Traffic Study*, Arthur D. Little, Inc., Final Report to DOT, Dec. 1979.

R. E. Dansby, "Welfare Economic Implications of Some Pricing, Capacity Investment and Advertising Decisions," unpublished doctoral dissertation, New York Univ. 1976.

______ **and R. D. Willig,** "Industry Performance Gradient Indexes" *Amer. Econ. Rev.*, June 1979, *69*, 249–60.

G. Iwata, "Measurement of Conjectural Variation in Oligopoly," *Econometrica*, Sept. 1974, *42*, 947–66.

R. Spady and A. F. Friedlaender, "Hedonic Costs and Economics of Scale in the Regulated Trucking Industry," *Bell J. Econ.*, Spring 1978, *9*, 159–79.

Trinc's Transportation Consultants, *Trinc's Bluebook of the Trucking Industry*, Washington 1977.

Price Distortions and Second Best Investment Rules in the Transportation Industries

By ANN F. FRIEDLAENDER*

There has recently been considerable interest in the question of second best pricing rules in the surface freight industries. Ronald Braeutigam first showed that if Ramsey pricing rules were followed to permit rail revenue to cover costs, rates should deviate from marginal cost in both the rail and trucking industries. Related studies by Richard Levin and Clifford Winston have analyzed the welfare costs of the existing and deregulated rate structures relative to the Ramsey rate structure.

This paper extends this second best analysis in the transportation industries to the problem of investment and shows that in the presence of price distortions, first best investment rules should generally not be used. While the specific second best investment rules that should be followed typically are complicated expressions that depend upon the underlying cost and demand functions, they generally imply that if the price distortions curtail output relative to its first best levels, offsetting investments should be made to increase capacity and output. Thus the findings of this paper resemble those in William Wheaton's analysis of urban highway investments.

This paper explores the relationship of investment rules to price distortions from both a theoretical and an empirical perspective. The theoretical analysis presents the argument in the simplest case of a single mode producing a single output and thus provides a relatively intuitive discussion of the problem. The empirical discussion then presents a simulation analysis of second best investment levels in the railroad industry in the context of intermodal competition and multiple outputs. The paper concludes with a brief discussion of the policy implications of the analysis.

I. Second Best Rules in a Single Output, Single Mode Framework

I assume that price and output changes in the transportation sector do not cause significant price changes throughout the economy and thus limit my analysis to the partial-equilibrium analysis of the transportation sector. Within the transportation sector, however, a number of distortions may exist. First, due to monopolistic pricing policies or regulation, prices may diverge from marginal cost; and second, the government may impose fuel or output taxes. The government's problem then is to choose the combination of taxes and investment levels that will maximize welfare given the existing distortions.

The welfare function is defined as the sum of consumers' surplus, producers' surplus, and government surplus. This is equivalent to the maximization of willingness-to-pay plus input tax revenues less the resource costs associated with transportation.[1] Thus the net benefit function can be expressed as

$$(1) \qquad B=\int_0^X q(z)dz-C(X,t,I)$$
$$-wI+tV(X,t,I)$$

*Massachusetts Institute of Technology. I am grateful to Subodh C. Mathur for research assistance he has provided in preparing this paper. Financial support from the National Science Foundation is gratefully acknowledged.

[1] The sum of consumers' surplus, producers' surplus, and net government surplus can be expressed as

$$B=\left[\int_0^X q(z)dz-qX\right]+[pX-C(X,t,I)]$$
$$+[uX+tV-wI]$$

where each of the bracketed terms represents the respective surplus and the variables have the same definitions as in the text. Simplifying and collecting terms yields equation (1).

where X represents the equilibrium industry output whose reduced form depends upon output taxes (u), fuel taxes (t), and the level of infrastructure (I), i.e., $X=X(u,t,I)$; q represents the price charged to consumers and is equal to the producers' price (p) plus the output tax (u), i.e., $q\equiv p+u$; $q(X)$ represents the inverse demand function; z represents a dummy of integration; $C(X,t,I)$ represents the short-run industry cost function;[2] wI represents the capital cost associated with the infrastructure; and $V(X,t,I)$ represents the conditional demand function for the taxed input, fuel.

The government has three policy instruments at its disposal: an output tax (u), a fuel tax (t), and the level of infrastructure (I), which is assumed to be publicly provided. Thus maximization of the benefit function with respect to these control variables yields the following first-order conditions:

$$\text{(2a)}\qquad B_u=(q-C_X)X_u+tV_xX_u=0$$

$$\text{(2b)}\qquad B_t=(q-C_X)X_t+tV_XX_t+tV_t=0$$

$$\text{(2c)}\qquad B_I=(q-C_X)X_I+tV_XX_I$$
$$+tV_I-C_I-w=0$$

where the subscripts denote differentiation with respect to the relevant arguments.[3]

The price-marginal cost differential is denoted by $q-C_X$. This can differ from zero because the government imposes output taxes or because of regulatory-induced or monopolistic price distortions. Thus $q-C_X\equiv u+m$, where u represents the output tax and m represents the exogenously determined price distortion.

In the absence of exogenous price distortions, inspection of equations (2a) and (2b) indicates that the optimal policy is to set both output taxes (u) and fuel taxes (t) equal to zero. If these taxes are zero, however, equation (2c) indicates that investment should be carried to the point where the marginal investment benefits, which are equal to the resource cost savings $(-C_I)$, equal the marginal investment costs (w). Let us denote by δ the difference between the marginal investment benefits and the marginal investment costs, i.e., $\delta\equiv-C_I-w$. Then if δ is positive, the marginal investment benefits are greater than the marginal investment costs and the infrastructure is curtailed relative to its first best levels. Conversely, $\delta<0$ implies that infrastructure is carried beyond its first best levels.

What is the relationship between price distortions, user charges, and investment rules? To see this, let us consider the case where an arbitrary fuel tax is employed and where an exogenous price distortion exists. In this case, the government's problem is to determine the optimal output taxes or subsidies and the optimal investment levels.

Inspection of equation (2a) indicates that in the presence of an arbitrary fuel tax the optimal output tax is a subsidy equal to the exogenous price distortion plus the marginal change in fuel tax revenues with respect to output, i.e.,

$$\text{(3a)}\qquad u=-(m+tV_X)$$

Thus the output tax should not only correct for any price distortions that may exist in

[2] Note that since the input prices are not assumed to change, I omit them in writing the cost function. I thus assume that the supply price of fuel is independent of the fuel tax or of fuel usage.

[3] To obtain the effects of changes in the policy instruments upon the equilibrium level of output, I implicitly differentiate the first-order conditions for a profit maximum, i.e., $\Pi_X\equiv(q-u)-C_X(X,t,I)=0$. (Note that I assume that the firm is a price taker; changing this assumption would add algebraic complexity without changing the results in a fundamental fashion.) From this differentiation it follows that

$$X_u=-\Pi_{Xu}/\Pi_{XX}=1/\Pi_{XX}<0$$

$$X_t=-\Pi_{Xt}/\Pi_{XX}=C_{Xt}/\Pi_{XX}$$
$$=C_{tX}/\Pi_{XX}=V_XX_u<0$$

$$X_I=-\Pi_{XI}/\Pi_{XX}=C_{XI}/\Pi_{XX}>0$$

Note that $\Pi_{XX}<0$ from the second-order condition for profit maximization and $C_{tX}=V_X$ from Shephard's lemma. Finally, it seems reasonable to assume that $C_{XI}<0$, i.e., an increase in infrastructure reduces marginal costs.

the transport industries, but should also correct for the revenue effect of the input price distortion. If u can be set in this quasi-optimal fashion, then equation (2c) indicates that investment should be carried to the point where the difference between marginal investment benefits and marginal investment costs just offsets the incremental loss in fuel tax revenues due to the change in investment, i.e.,

$$\delta = -tV_I \tag{3b}$$

Since infrastructure I represents the fixed factor, it follows that $V_I < 0$; an increase in infrastructure should lead to a reduction in the amount of the variable factors utilized. Thus equation (3b) indicates that in the presence of an arbitrary fuel tax and a quasi-optimal output tax, $\delta > 0$, and investment should be curtailed relative to its first best levels. This result makes intuitive sense since additional investment causes a substitution against fuel and hence a reduction in fuel tax revenues and benefits. To counteract this effect, investment should be curtailed.

This suggests that if institutional constraints exist to prevent a subsidy (as they apparently do), infrastructure should be expanded relative to its quasi-optimal levels. In this case, the divergence between marginal investment benefits and costs should equal the value of the disallowed subsidy less the loss in fuel tax revenues arising from the change in the infrastructure, i.e.,

$$\delta = -mX_I - t(V_X X_I + V_I) \tag{4}$$

The first term of this expression is unambiguously negative, while the second term reflects two conflicting pressures: the revenue loss due to the substitution against fuel (tV_I), which I call the substitution effect; and the revenue increase due to the increase in capacity and usage ($tV_X X_I$), which I call the output effect. If the output effect is greater than the substitution effect, the second term is also unambiguously negative. Hence $\delta < 0$, indicating that infrastructure should be expanded relative to its first best levels. If, however, the substitution effect is greater than the output effect, the sign of the second term is ambiguous, indicating that $\delta \lesseqgtr 0$. Hence whether the infrastructure should be greater or less than its first best levels cannot be determined a priori, and depends upon the relative magnitudes of the price distortions and the net revenue loss in fuel taxes arising from the investment. Nevertheless, a comparison of equations (4) and (3b) clearly indicates that in the absence of a quasi-optimal output subsidy, investment should be expanded relative to the levels that would exist with such a subsidy. This makes intuitive sense since the increase in capacity stimulates increased output and hence operates in the same fashion as a subsidy.

II. Second Best Investment in the Context of Rail-Truck Competition

While useful in highlighting the problem of determining quasi-optimal investment rules in the presence of price distortions, the above analysis is somewhat artificial in that it only assumes a single mode and a single output. In fact, however, each mode typically carries a wide range of different kinds of freight and there is a high degree of substitutability among the various modes and interdependencies among their demand functions. Thus price distortions in one mode will typically imply that offsetting price distortions should be made in other modes (for example, see Braeutigam). Consequently, a full analysis of second best investment rules requires a multi-commodity, multi-output framework.

Unfortunately, however, data are lacking to undertake a full analysis of the multimodal multiple-output relationships in the transportation industries. Nevertheless, an initial approximation can be made by utilizing the cost and demand functions I developed with Richard Spady. We estimated the marginal cost and demand functions for manufactured and bulk commodities in the rail and truckload trucking industries in 1972 for the Interstate Commerce Commission's Official Territory, which comprises the New England, Middle Atlantic, and Central

TABLE 1—NET BENEFITS, INFRASTRUCTURE, OUTPUT AND RATES, VARIOUS SCENARIOS, BY RAIL AND TRUCK, OFFICIAL TERRITORY, 1972

			Distorted Equilibrium	
	Historical Equilibrium	Competitive Equilibrium	Historical Distortions	30 Percent Distortions
Net benefits ($ bil.)	5.7	6.9	6.3	6.6
Capital stock ($ bil.)	19.0	8.8	7.0	9.2
Output (bil. ton-miles)				
Rail, manufactured	83.1	30.3	42.0	12.9
Rail, bulk	57.1	76.5	43.6	54.2
Truck, manufactured	27.5	49.2	33.2	35.0
Truck, bulk	13.4	22.4	14.8	14.1
Rates (¢/ton-mile)				
Rail, manufactured	2.4	4.6	4.0	9.3
Rail, bulk	2.1	1.5	2.6	2.2
Truck, manufactured	6.1	4.5	5.9	6.8
Truck, bulk	5.9	3.7	5.8	5.7
$(P-MC)/P$				
Rail, manufactured	−0.628	0.0	−0.628	.30
Rail, bulk	0.074	0.0	0.074	.30
Truck, manufactured	0.192	0.0	0.192	.30
Truck, bulk	0.300	0.0	0.300	.30

States.[4] Although our analysis is based on stylized or composite firms in each industry, it provides a useful basis to compare the benefits under a distorted and competitive equilibrium.

To analyze the significance of the second best investment rules, I analyze how benefits are affected by following different investment criteria in the railroad industry. While it would obviously be useful to consider the question of highway infrastructure, data are currently lacking to perform this analysis.[5] I thus assume that the goal of the public authority is to maximize net benefits with respect to railroad infrastructure, given existing price distortions and taxes, and given the estimated rail and trucking cost and demand functions.

The significance of following second best investment rules can be determined by estimating the net benefits under four different scenarios: 1) the existing equilibrium which assumes historical (1972) price distortions and infrastructure levels; 2) the long-run competitive equilibrium which assumes no price distortion and first best investment rules; 3) the second best equilibrium which assumes historical price distortions and second best investment rules; and 4) the second best equilibrium which assumes a 30 percent distortion in all markets and second best investment rules.

Table 1 summarizes the relevant information and indicates that the historical equilibrium is far from a competitive equilibrium. Not only is the infrastructure more than twice what it would be in a long-run competitive equilibrium, but there are significant price distortions, with manufactured goods receiving substantial subsidies in the rail industry, and other commodities having price-marginal cost ratios ranging from 7 to 30 percent. Thus a movement to a long-run competitive equilibrium would cause substantial movements in rates and outputs, with rail rates for manufactured goods rising substantially, other rates falling substantially, and manufactured traffic moving from rail to truck. As expected, benefits rise in moving from the historical equilibrium to a long-run competitive equi-

[4] See our study for a full discussion of the derivation of the relevant cost and demand functions and the simulation analysis used to obtain the equilibrium.

[5] Although the trucking cost functions reflect the long-run costs facing the firm, they abstract from the role of highway infrastructure and thus do not reflect the full costs of highway transportation. Efforts are currently underway to introduce highway infrastructure into the analysis.

librium, but their increase is relatively modest in comparison to the changes in rates and outputs.

The significance of following second best investment rules can be seen by comparing the distorted equilibria to the long-run competitive equilibrium. If the historical distortions are assumed to be maintained, railroad infrastructure should be curtailed relative to its first best levels. This makes intuitive sense since the rail subsidy to manufactured goods is significantly higher than the implicit rail tax on bulk commodities. Hence output should be curtailed through infrastructure investments to offset the inefficiencies of the regulated pricing structure. In contrast, infrastructure should be expanded if all sectors have a 30 percent markup over marginal costs. In this case, output is uniformly curtailed by the pricing structure, and infrastructure should be expanded to offset this effect. However, the changes in infrastructure and in net benefits appear to be relatively modest in comparison to the rate and output changes that occur in each of these scenarios. This implies that the optimal level of infrastructure may be relatively insensitive to the choice of the investment rules and that the net benefit function may be relatively flat.

III. Policy Implications

The policy results of this analysis are both discouraging and encouraging. The discouraging element arises from the analytic complexity and lack of straightforward second best investment rules. Thus the theoretical analysis indicates that the presence of price distortions makes the evaluation of investments fundamentally more complex. Although it would in principle be possible to develop cost and demand functions that would enable the analyst to determine second best investment rules, in practice the complexities implied by second best investment rules would tax already overextended bureaucracies.

Fortunately, however, the empirical analysis indicates that the quantitative implications of the second best investment rules may not be that substantial. If, in fact, the benefit function is rather flat over a wide range of distortions and different levels of investment and if the quasi-optimal investment levels are relatively constant, then relatively little social cost results from following first best investment rules in the presence of substantial price distortions.

Ultimately, however, the second best investment rules depend upon the nature of the distortions and the nature of the underlying cost and demand functions. Thus just as relatively little can be definitively inferred from a theoretical analysis of second best investment rules, relatively little can be inferred from the particular cases considered in this paper. Nevertheless, the findings of this paper indicate that the costs of following wrong investment rules may be relatively slight. If corroborated by subsequent analysis, this will be an encouraging finding since it is clear that the policy analysis associated with first best investment rules is considerably less complex than that associated with second best investment rules.

REFERENCES

R. R. Braeutigam, "Optimal Pricing with Intermodal Competition," *Amer. Econ. Rev.*, Mar. 1979, *69*, 38–49.

Ann F. Friedlaender and Richard H. Spady, *Freight Transport Regulation: Equity, Efficiency and Competition in the Rail and Trucking Industries*, MIT Press: Cambridge, Mass. 1981.

R. Levin, "Railroad Regulation, Deregulation, and Workable Competition," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 394–98.

W. C. Wheaton, "Price-Induced Distortion in Urban Highway Investment," *Bell J. Econ.*, Autumn 1978, *9*, 622–32.

C. Winston, "Welfare Effects of ICC Rate Regulation Revisited," Center for Transportation Studies, MIT, 1980.

Railroad Regulation, Deregulation, and Workable Competition

By RICHARD C. LEVIN*

In documenting the shortcomings of regulation, economists have typically compared the policy under scrutiny to an ideal social optimum. But, in the current enthusiasm for deregulation, few have paused to point out that in many regulated industries deregulation is unlikely to achieve optimality. In the case of railroads, indivisibilities, pervasive economies of joint production, and high costs of entry lead inevitably to small-numbers competition under deregulation and consequently to the likely persistence of prices in excess of marginal cost. Under such conditions, it is appropriate for the policy analyst to move beyond the comparison of actual regulatory performance with an ideal regime to a focus on two hitherto neglected questions: Will complete deregulation improve welfare relative to the regulated status quo?, and Are there simple and workable regulatory policies that dominate both the status quo and unregulated outcomes? In this paper I examine the evidence bearing on these questions in the railroad case.

These questions seem particularly appropriate as we enter an era of substantial regulatory reform at the Interstate Commerce Commission. The Railroad Revitalization and Regulatory Reform Act of 1976 (the 4R Act) removed maximum rate regulation except in those cases where a railroad is determined by the ICC to possess "market dominance." This potentially sweeping reform was scuttled by the ICC, which established criteria for market dominance so stringent as to subvert the legislative intention to introduce significant price flexibility. While the ICC subsequently moved under new leadership to weaken the market dominance standard, congressional debate on additional regulatory reform legislation focused on design of a mechanism to protect "captive shippers" from the undue exercise of railroad market power. To oversimplify slightly the Staggers Rail Act of 1980, all rates below a critical ratio of revenue to variable cost will be *per se* legal; other rates will be governed by a rule of reason. In the analysis that follows I will compare the welfare consequences of the old regime, full deregulation, and the imposition of both high and low ceilings on the ratio of revenue to variable cost.

I. Simulating Alternative Regulatory Policies

I examine the consequences of full and partial deregulation using a simulation model developed with the use of demand equations drawn from the recent empirical literature and cost estimates from the ICC. Full details of the simulation model are discussed in my forthcoming article. To summarize briefly, railroad industry demand curves for manufactured commodities are derived from the multinomial logit model of modal choice estimated in my 1978 article. The logit model assumes, somewhat unrealistically, that the total size of the transport market is fixed, but it should be noted that the point elasticities of rail demand at observed prices derived using the logit approach fall in the same range as those estimated using less restrictive assumptions by Ann Friedlaender and Richard Spady (1981), Tae Hoon Oum, and Clifford Winston (1979). For the major bulk commodities (coal, metallic ores, and field crops), I use modified versions of the demand equations estimated by Friedlaender and Spady (1978), and Winston's (1978) results are used

*Yale University. I am indebted to Richard Kolsky for computational assistance. This research was supported by grants from the Eno Foundation and the Alfred P. Sloan Foundation.

for produce.

Market equilibria are simulated under a variety of assumptions regarding truck prices, rail costs, and the degree of interrailroad competition. In this paper, I report results under the following alternative assumptions: 1) truck prices either remain unchanged or fall 10 percent with deregulation and 2) railroad marginal costs (= average variable costs) either remain unchanged or fall 10 percent. These two pairs of assumptions produce four possible scenarios. For each, I consider the full range of values of the degree of interrailroad (intramodal) competition, which is parameterized as the ratio of the demand elasticity perceived by a representative firm and the demand elasticity of the rail industry. The relative elasticity parameter varies from a value of one in the case of monopoly to values approaching infinity when the number of competitors is large or behavior is highly rivalistic. In a symmetric Cournot model, the relative elasticity parameter is simply equal to the number of competitors, but behavior which is more or less rivalistic than under Cournot assumptions will generate relative elasticities which are, respectively, larger or smaller than the number of firms in the market.

Three changes in regulatory policy are simulated: 1) complete deregulation of railroad rates; 2) complete rate flexibility up to a ceiling price equal to 250 percent of average variable cost; and 3) flexibility up to a ceiling of 160 percent variable cost. The lower ceiling corresponds to ratio of revenue to variable cost which the ICC held to be presumptive of market dominance in its initial interpretation of the 4R Act. Under the 1980 legislation, rates below this level will be per se legal.

The effects of these policy changes are measured against two benchmarks: the regulated status quo and a regime of ideal or, more precisely, second best optimum regulation. While it would be desirable in principle to use very recent data to calculate the welfare consequences of the regulated status quo, I have used the year 1972 as a benchmark for two reasons. First, when this work was first undertaken the 1972 Census of Transportation was the most recent available source of data on rail and truck market shares for manufactured commodities. Second, data from the most recent census (1977) may already show some effects of movement toward deregulation. Hence, the results reported here should be strictly interpreted as the predicted consequences of full and partial deregulation if these policies had been implemented in 1972. Changes in relative costs, rates, and service qualtiy of railroads and competing transport modes in the ensuing years do not permit a clear inference about the direction of bias in predictions based on 1972 data.

The indivisibilities, jointness, and fixed costs which make small numbers competition an inevitable consequence of deregulation also render the traditional measure of static deadweight loss incomplete as a welfare indicator. A regime of marginal cost pricing would eliminate the deadweight loss. But marginal cost pricing is not only implausible as the outcome of deregulation, it is also a questionable regulatory objective, since the railroads would incur substantial losses. Without subsidy, reduction of the short-run welfare loss to zero would cause the long-run deterioration of the industry's capital stock. Under such conditions, if consideration is limited to solutions involving private ownership, a more plausible regulatory objective is to minimize static deadweight loss subject to the achievement of a normal rate of return.

Optimal regulation in this sense involves a variant of Ramsey pricing suggested by Ronald Braeutigam in which rail prices deviate from marginal costs, but truck prices are competitively determined. I set the rate of return constraint at 8 percent, roughly equal to the cost of railroad capital in 1972. The railroad capital stock is valued at 90 percent of reproduction cost as estimated by Douglas Caves, Laurits Christensen, and Joseph Swanson. I assume a 10 percent reduction in the capital stock as a conservative estimate of the amount of uneconomic excess line capacity in the present *U.S.* railroad system.

TABLE 1—RATES OF RETURN AND WELFARE LOSSES UNDER ALTERNATIVE REGULATORY REGIMES

Competitive conditions			Regulatory Regime													
Rail costs	Truck rates	Degree of competition	Base case		Ramsey prices		Deregulation			Price<2.5 x var.cost			Price<1.6 x var.cost		
			ROR (%)	DWLR ($mill.)	ROR (%)	DWL* ($mill.)	ROR (%)	DWL/DWLR	DWL/DWL*	ROR (%)	DWL/DWLR	DWL/DWL*	ROR (%)	DWL/DWLR	DWL/DWL*
No change	No change	1	1.95	742	8.00	1058	11.57	8.18	5.74	10.11	4.42	3.10	6.45	1.19	0.84
		3					8.45	1.67	1.17	8.28	1.59	1.12	6.13	0.85	0.60
		5					6.30	0.73	0.51	6.28	0.72	0.51	5.38	0.53	0.37
	Decline 10%	1				1275	10.71	7.79	4.53	9.43	4.30	2.50	6.11	1.19	0.69
		3					7.76	1.56	0.91	7.62	1.49	0.87	5.76	0.82	0.48
		5					5.76	0.67	0.39	5.74	0.67	0.39	5.04	0.51	0.30
Decline 10%	No change	1		2012		933	12.22	3.15	6.78	10.30	1.55	3.34	6.29	0.36	0.79
		3					8.96	0.65	1.40	8.67	0.60	1.29	6.12	0.29	0.63
		5					6.70	0.28	0.61	6.66	0.28	0.60	5.45	0.19	0.41
	Decline 10%	1				1120	11.33	3.00	5.39	9.64	1.52	2.72	5.98	0.37	0.66
		3					8.24	0.61	1.09	8.03	0.57	1.03	5.79	0.29	0.52
		5					6.12	0.26	0.47	6.09	0.26	0.47	5.11	0.18	0.33

Notes: *ROR*: Rate of return; *DWL*: Deadweight loss under indicated scenario; *DWLR*: Deadweight loss under regulation; *DWL*: Deadweight loss with Ramsey Prices.

II. Rates of Return and Welfare Losses under Alternative Regimes

Table 1 reports the results of the simulation experiments. In the 1972 regulated base case, the railroad industry achieved a rate of return (adjusted for appropriate capitalization of roadway maintenance expenditures) of 1.95 percent, a figure roughly equal to the average annual return over the period 1968–77. The size of the deadweight loss depends on whether it is assumed that observed marginal costs have been elevated as a consequence of regulation. If not, the loss is simply a summation of the familiar welfare triangles across the 353 markets included in my sample, appropriately scaled in accordance with the sample coverage. If, however, it can be assumed that marginal costs will fall 10 percent with deregulation, a trapezodial area representing the cost savings on existing output plus an increment to the conventionally measured triangle must be added to arrive at the total static welfare cost of regulation.

The welfare loss associated with a Ramsey optimum also depends on the assumed behavior of railroad costs with deregulation. Lower marginal costs permit the rate of return constraint to be satisfied at a smaller total welfare cost. On the other hand, if truck rates decline 10 percent with deregulation the demand for rail services is reduced and larger deviations of price from marginal cost are needed to satisfy the rate of return constraint. Under any combination of these assumptions about rail costs and truck rates, a Ramsey price regime would generate a static deadweight loss on the order of $1 billion annually as the price of achieving a financially viable privately owned railroad industry.

The most striking result obtained in simulating the outcome of complete deregulation is the importance of the degree of interrailroad competition. Predicted rates of return and welfare losses vary only moderately in response to plausible reductions in rail costs and truck rates, but a modest degree of interrailroad competition has profound effects relative to the pure monopoly case. As Table 1 indicates, rates of return predicted on the assumption that each railroad market is served by a monopolist or perfectly collusive oligopolists range from 10.71 to 12.22 percent, well in excess of the required 8 percent. Yet when the elasticity of firm demand is assumed to be three times the industry elasticity, the rate of return approximates the cost of capital. Even more strikingly, the welfare loss falls by a factor of five as the relative elasticity parameter increases from one to three.

A second result of significant interest is that the welfare consequences of ceiling price regulation depend very much on the level of the ceiling. A maximum ratio of revenue to variable cost of 1.6 would apparently be inadequate to generate a normal rate of return even in the absence of interrailroad competition. Such a stringent ratio test would produce some reduction in deadweight loss relative to the status quo, but it would fail to provide a return adequate to assure the long-run viability of the industry. Moreover, the low ceiling constrains not only true rail monopolists, but it is binding in many markets where interrailroad competition is present. The more generous ceiling examined here seems to have more favorable consequences. In monopoly markets, a 250 percent ceiling cuts by nearly half the welfare loss that would obtain under complete deregulation, but profits fall by only 12–16 percent, preserving financial viability. On the other hand, in the presence of a moderate degree of interrailroad competition the higher ceiling would prove a binding constraint on only a small fraction of railroad traffic.

Though the virtues of ceiling price regulation in truly monopolized markets are apparent, the gains from introduction of an appropriate degree of interrailroad competition are even more dramatic. Since both Ramsey and conjectural variation equilibria involve proportionate reductions from the output at which price equals marginal cost, there will necessarily be some degree of competition which produces an outcome identical to the Ramsey solution. In the railroad industry, this ideal ratio of firm to industry demand elasticity is in the neighborhood of three. Of course, excessive rivalry among railroads can lead to subnormal rates of return and consequent long-run welfare losses, but my calculations indicate that a relative elasticity in excess of ten is required to drive the deregulated rate of return below that earned in the regulated base case. Since interrailroad competition most commonly involves two or three firms, such a high degree of rivalry seems implausible in most markets.

These findings taken together point toward the conclusion that deregulation in the presence of a moderate degree of interrailroad competition will almost certainly enhance social welfare. The tradeoff between short- and long-run efficiency may even be such that pure monopoly is preferable to base case regulation. Indeed, this is not improbable since it has been widely argued that persistent deficiency in the return to capital has impeded railroad capital formation, raised operating costs, reduced the quality of service and stifled innovation. Nevertheless, society's choice is fortunately not between status quo regulation and complete monopoly, since truly "captive" shippers generate only a modest fraction of rail traffic. In reality, a workable degree of interrailroad competition is probably present in most markets for transport of agricultural and manufactured commodities. Moreover, interrailroad competition will surely increase as a consequence of recent measures which restrict the power of rate bureaus, remove protective traffic conditions, and grant considerable freedom to shippers and carriers to negotiate contract rates.

III. Policy Implications

The evidence presented here suggests that the two questions posed in the introduction to this paper may be answered in the affirmative. Complete deregulation of railroad rates is likely to improve welfare relative to the regulated status quo, and there appear to be simple and workable regulatory policies that dominate both the status quo and unregulated outcomes.

The simplest such policy alternative is the imposition of a price ceiling at a specified percentage of variable cost. While a ratio test is simple in principle, there is significant danger of maladministration of even so simple a rule. The accuracy of the cost calculations admissable in regulatory proceedings becomes crucial to proper enforcement, and the past performance of the ICC in developing costing methodology has been dismal at best. Moreover, as the evidence in this paper indicates, too stringent a ratio test will perversely constrain behavior in workably competitive markets and prevent the industry from earning normal returns. These considerations suggest that a price ceiling, a potentially valuable instrument in reducing deadweight loss in truly monopolistic markets, should be set at a rather high level, much closer to 250 percent of variable cost than to the 160 percent rule established under the 4R Act.

The new deregulation legislation is sufficiently flexible to permit this outcome, since the low ratio it specifies only places a ceiling on the per se legality of rates. At a second ceiling, which may be set as high as 200 percent after four years, the burden of proof of reasonableness shifts from protestant to proposing carrier. A lenient interpretation of the standards for reasonableness could permit the realization of substantial social gains from residual rate regulation.

Whatever the benefits of ceiling price regulation, they are overshadowed by the benefits of workable interrailroad competition. The current wave of merger applications promises to produce a major restructuring of the *U.S.* railroad network. In deciding these cases, the ICC will have an opportunity to play a constructive role in preserving interrailroad competition where it is threatened. In its decision in the Burlington-Frisco case, the newly constituted ICC has already revealed its sympathy for end-to-end consolidations, which usually, although not always, have minimal anticompetitive consequences. Forthcoming merger cases, however, will involve railroads which compete directly in numerous markets. In such cases, it may be possible to preserve competition by a variety of instruments short of prohibiting merger, such as selling trackage rights or one of two parallel lines to a third party, or requiring reciprocal switching agreements with competitors. Realizing the potential benefits of rate deregulation is inextricably linked to careful guidance of the ongoing process of restructuring the railroad network.

REFERENCES

R. Braeutigam, "Optimal Pricing with Intermodal Competition," *Amer. Econ. Rev.*, Mar. 1979, *69*, 38–49.

D. Caves, L. Christensen, and J. Swanson, "Productivity in U.S. Railroads, 1951–1974," *Bell J. Econ.*, Spring 1980, *11*, 166–81.

Ann Friedlaender and Richard Spady, "Derived Demand Functions for Freight Transportation," in Ann Friedlaender, ed., *Alternative Scenarios for Federal Transportation Policy: Second Year Report*, Vol. I, Cambridge, Mass., Feb. 1978.

______**and**______ *Freight Transport Regulation*, Cambridge, Mass., 1981.

R. Levin, "Allocation in Surface Freight Transportation: Does Rate Regulation Matter?," *Bell J. Econ.*, Spring 1978, *9*, 18–45.

______, "Railroad Rates, Profitability Welfare under Deregulation," *Bell J. Econ.*, Spring 1981, forthcoming.

T. Oum, "A Cross-Sectional Study of Freight Transport Demand and Rail-Truck Competition in Canada," *Bell J. Econ.*, Autumn 1979, *10*, 463–82.

C. Winston, "Mode Choice in Freight Transportation," working paper, Univ. California-Berkeley, Nov. 1978.

______, "A Disaggregate Model of the Demand for Intercity Freight Transportation," working paper, M.I.T., Aug. 1979.

Interstate Commerce Commission, *Rail Carload Cost Scales, 1972,* Washington 1974.

Information and the Law: Evaluating Legal Restrictions on Competitive Contracts

By JANUSZ ORDOVER AND ANDREW WEISS*

In markets with many buyers and sellers, some of whom are imperfectly informed, legal restrictions on the terms of contracts may improve allocative efficiency. We examine three quite different forms of government intervention and show that in each case interference with what may appear to be "competitive" market outcomes may improve the allocation of resources. The contractual restrictions we discuss are: a) forbidding banks from denying loans to an entire class of borrowers; b) forbidding stores from charging prices which are high relative to the average price in the market; c) forbidding contract provisions which impose some of the cost of product failure on consumers.

I. Exclusion of Borrowers

We should not be surprised to find banks charging different borrowers different interest rates. What is surprising is that banks, at times when the usury laws do not seem to be binding, may refuse to lend to an entire class of borrowers at any interest rate. The fact that banks are not using the price system (interest rates) to allocate credit, i.e., giving loans to those borrowers willing to pay the highest interest rate, suggests that the credit market is not operating in the way we would expect from the usual analysis of competitive markets with perfect information. Before showing why forbidding banks from excluding borrowers can improve welfare, we need to understand why they will practice this form of exclusion.

If usury laws are not binding, a profit-maximizing bank which excludes a group of borrowers from credit must have determined that there were no interest rates or collateral requirements at which loans to those borrowers would be profitable. That is, as interest rates go to infinity it cannot be the case that the return to the bank also goes to infinity, even if the bank could also change collateral requirements with the increase in interest rates. (For a rigorous analysis of market equilibria with credit rationing see Joseph Stiglitz and Weiss, 1980, 1981.)

There are two reasons why bank profits may not rise with increases in interest rate. First, borrowers will tend to favor projects with a higher probability of default when the interest rate is increased. This is because the interest rate is not paid in very bad states—when the borrower defaults; thus the increase in the interest rate has a greater impact on projects with a low probability of default. If a borrower were initially indifferent between two projects at interest rate r^* then at interest rate $r^* + \varepsilon$ he would undertake the project with the higher probability of default. This incentive effect of interest rates would discourage banks from raising the interest rate when faced with an excess demand for loanable funds.

Second, borrowers who are deterred from borrowing by the high cost of capital may be precisely the borrowers to whom the bank could most profitably lend money. Let us assume that the bank has prescreened borrowers, so that within each class of borrowers the expected total return on a loan is constant, and further assume that each borrower has only one available project. These projects can be ranked so that a "riskier" project is a mean preserving spread of a "safer" one. In Figure 1 we show the return

*New York University and Bell Laboratories, respectively. The views here are our own, and not the Bell System. Ordover's research was financed by a grant from the National Science Foundation.

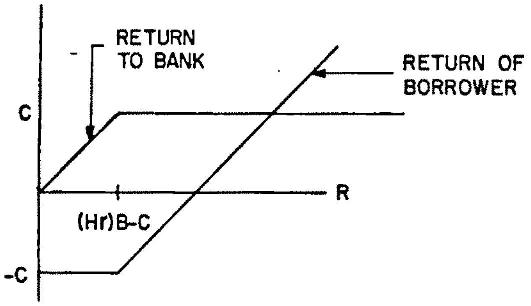


FIGURE 1

to the bank and to the borrower as a function of the realized total return on the project R, the amount borrowed B, the interest rate r, and the collateral demanded C. From Jensen's inequality and the convexity of the borrower's return function, we can see that borrowers make higher profits on riskier projects; on the other hand, a risk-neutral bank prefers to finance safer projects. Therefore, if borrowers are risk neutral, the marginal borrowers, the ones deterred from borrowing by a higher interest rate, are those who would have invested in safe projects, which are the most profitable loans for the bank.

The intuition behind this result is that a riskier project has higher returns in good states and lower returns in poor ones. The higher returns in the good states benefit the borrower who gets to keep all the returns on the project above the cost of the loan, but do not benefit the bank which simply repaid the loan with interest regardless of the extent by which the actual return exceeds this commitment. The lower return in poor states do not hurt the borrower who defaults on loans in those states, but do hurt the bank which gets all the assets of the borrower in the case of a default. The lower return implies less assets available to the bank. Therefore, the expected profit of a borrower is an increasing function of the riskiness of his project while the expected profit of the bank decrease with the riskiness of the project. Higher interest rates therefore discourage the safe borrowers whom the bank preferred. This sorting effect, in general, reinforces the incentive effect of interest rates in discouraging banks from raising their interest rate in response to an excess demand for credit.

Stiglitz and Weiss (1981) also show that collateral requirements may have similar adverse selection effects so that increasing collateral requirements in response to an excess demand for credit may also decrease the banks profits. The most striking of the sorting effects of collateral requirements is that if borrowers have decreasing absolute risk aversion (an assumption with overwhelming empirical support) then raising collateral requirements discourages those borrowers who would have intested in the safest projects. Higher collateral requirements also results in smaller scale projects which may (if there are increasing returns to scale in investment) lower bank profits; and the ability to meet the collateral requirements of banks may be due to the high returns on previous risky investments so that those borrowers able to satisfy high collateral requirements are precisely those borrowers with the highest preference for risky investments.

In Figure 2 we draw the return to a bank from lending to different groups at different interest rates (holding collateral requirements fixed). We are assuming that the combined incentive and sorting effects of interest rates will cause the return to a bank to decline with increases in the interest rate it charges borrowers when that interest rate is sufficiently high.

Figure 2 shows a situation in which at a cost of funds to the bank of i^*, type a borrowers receive loans at interest rate r_a and type b borrowers get credit at interest rate r_b. Type c and d borrowers are totally excluded from the credit market because the maximum return a bank could get from lending to either c or d is less than i^*, the competitive return to depositors. The maximum return to a bank for a class of borrowers is a function of the expected value of the projects they undertake, the riskiness of the set of feasible projects available to the borrower, the heterogeneity of the borrowers, and the degree to which a bank can monitor the actions of borrowers. (The importance of the last two effects can be seen by noting that if borrowers were homogeneous and banks could monitor investments

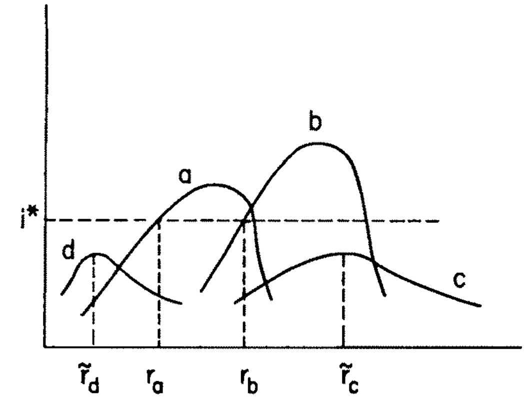


FIGURE 2

the return for the bank would not be affected by the adverse sorting and incentive effects noted above.) Therefore, groups of borrowers may be denied loans even if the total return to loans to group d at $\tilde{r}_d$ and group c at $\tilde{r}_c$ is higher than the expected total returns on the investments groups a and b are making. However, because the former group is less well known to the bank so that its actions can't be as accurately monitored and cannot be as readily sorted into different risk types, or because the high total return projects of the excluded borrowers represent riskier projects and so have lower returns to the bank, they are excluded from the credit market. In this case, forcing banks to lend to all borrowers at some interest rate would increase the expected total return per dollar loaned. Types c and d borrowers would get credit at interest rates $\tilde{r}_c$ and $\tilde{r}_d$, respectively, while the interest rates for a and b borrowers would rise. By also imposing proportional taxes on the profits of borrowers, and transfers to depositors, the government could ensure that the supply of loanable funds was unchanged by its intervention, and hence that the aggregate return on investments was increased because of the "better" allocation of credit.

II. Forbidding Unconscionably High Prices

The very existence of a price distribution of a homogeneous product suggests again that the market is not acting in the way we would expect if all traders were perfectly informed. It seems reasonable to assume that consumers who purchase at an unconscionably high price and later sue to get the sale negated were uninformed when they made the purchase.

In order to analyze the impact of government restrictions on relative prices, one needs to formulate a model in which there is substantial dispersion in the prices of a homogeneous product. Here we borrow from models developed independently by Hal Varian and Robert Rosenthal. Using the Varian notation, there are M uninformed and I informed consumers, each of which buys one and only one unit of the commodity at prices less than or equal to below r and no units at prices above r (M, I, and r are determined exogenously). The *informed* consumers know all the prices charged in the economy and buy at the lowest price store. The *uninformed* consumers buy at the first store they enter, charging a price $p \leqslant r$. Free entry determines the number of (identical) stores n and ensures that each store makes zero profits. Varian also assumes that a store must charge the same price to all its customers and that for each store average costs are decreasing. There is a unique mixed strategy symmetric equilibrium in which the distribution of prices charged by each store, $F(p)$ is defined by

$$(1)\quad F(p)=$$

$$\begin{bmatrix} 1 & p \geqslant r \\ 1-\left[\dfrac{\pi_f(p)}{\pi_f(p)-\pi_s(p)}\right]^{\frac{1}{n-1}} & p^* < p < r \\ 0 & p \leqslant p^* \end{bmatrix}$$

where p^* is the average cost of a store selling to $I+M/n$ customers, $\pi_f(p)$ is the profit of a firm charging price p when it fails to be the lowest priced store, and $\pi_s(p)$ is the firm's profit when it is the lowest priced store. In equilibrium $\pi_f(p)<0$ for all $p<r$.

Let us assume that each store has a fixed cost k and constant marginal cost c. Then

$$(2)\qquad \pi_s(p)=(p-c)\left[I+\frac{M}{n}\right]-k$$

and

$$(3) \qquad \pi_f(p)=(p-c)\frac{M}{n}-k$$

From the equilibrium condition that $\pi_f(r)=0$, $n=(r-c)M/k$. Substituting into equation (1), and letting $p-c=\tilde{p}$ and $r-c=\tilde{r}$, the distribution of lowest prices is

(4)

$$1-[1-F(p)]^n=1-\left[\frac{k}{I}\left(\frac{1}{\tilde{p}}-\frac{1}{\tilde{r}}\right)\right]^{\frac{\tilde{r}M}{\tilde{r}M-k}}$$

Eliminating high prices has the same impact in this model as lowering r, whether or not the high price is defined in absolute terms or relative to the lowest price offered, p^*. Therefore, we can find the effect on the equilibrium price distribution of court decisions penalizing high prices by differentiating $[1-F(p)]^n$ with respect to r. Let

$$[1-F(p)]^n=A \text{ and } \frac{k}{I}\left[\frac{1}{\tilde{p}}-\frac{1}{\tilde{r}}\right]=B$$

then

(5)

$$\frac{dA}{d\tilde{r}}=A\left[\frac{-Mk}{[\tilde{r}M-k]^2}\ln B+\frac{kM}{I\tilde{r}(\tilde{r}M-k)B}\right]$$

Since $B>0$ the second term in the brackets is positive. The first term is nonnegative for all values of p for which $B \leqslant 1$. However, B takes on its largest value where $\tilde{p}$ is at its smallest value, which is at $\pi_s(p)=0$. Substituting for $M/n=k/\tilde{r}$ in equation (2), and p^*-c for $\tilde{p}$ in the definition of B, we see that $B \leqslant 1$, therefore $d(1-F(p))^n/dr>0$ and $d[1-(1-F(p))^n]/dr<0$. Thus, decreasing the maximum allowed price (lowering r) will cause a lower distribution (by first order stochastic dominance) of the lowest price. From equations (1), (2), and (3), it is straightforward to show that $dF(p)/dr<0$. Consequently both informed and uninformed consumers benefit from an unconscionability rule. Since we've assumed free entry (zero profits) the profits of firms are unchanged; therefore, this is a Pareto-improving change.

This result uses the assumption that there are consumers with zero search costs. Avishay Braverman has shown that the nature of equilibrium in models similar to Varian's can be sensitive to the distribution of search costs.

Suppose that all individuals have positive costs of becoming informed and that an uninformed individual knows neither the distribution of prices nor the price charged by any store. Therefore, the decision to become informed is a function of the individual's expectations about the distribution of prices and the individual's own cost of information. In game theoretic terms, the strategy of an individual is a decision to become informed and buy at the lowest price store, or stay uninformed and buy at some randomly chosen store. The strategy of a firm is the choice of a price or price distribution to offer. Let us assume that individuals believe that the unconscionability rule will cause each firm to offer the same price; then no consumer will become informed, and each store will charge the monopoly price r. This is a Nash equilibrium; no consumer or firm would be better off by changing its strategy. It is also apparent that the expectations of consumers were rational: after the unconscionability rule was imposed the distribution of prices did collapse to a single price.

Even if all consumers eventually learn the prices charged by all firms (albeit after having made their purchase) and could invoke the unconscionability rule to revoke their sale and switch to the lowest price store, the equilibrium described above may still be stable. Remember that the unconscionability rule nullifies contracts only if the price is unreasonably higher than the *average* price. If there are many stores, then one deviant store charging a low price will have a very small effect on the average price. Therefore, the welfare implications of an unconscionability rule are quite sensitive to its effect on expectations, and to the precise information structure of the market.

III. Forcing Firms to Assume the Full Cost of Product Failure

It is commonly agreed that when markets are perfectly competitive, in particular: consumers are fully informed about the risk of a product, producers costlessly distinguish between the various types of buyers of their product, and contract terms do not affect the behavior of buyers, then the choice between consumer liability and strict producer liability is immaterial from an efficiency standpoint. This result does not obtain, however, when consumers differ according to unobservable innate carefulness (see Ordover for a full exposition of this model).

Let us assume that there are only two types of consumers in the population, careful and careless ones, that the product-cum-warranty market is perfectly competitive, and that the probability of product failure depends only on the (known) quality of the product and on the unobservable characteristics of the consumer. By a simple extension of the work of Michael Rothschild and Stiglitz, one can show that, in the absence of government restrictions on contract terms, if an equilibrium exists, it will be characterized by careful consumers signaling their characteristic by assuming some of the risk of product failure. The careless consumers prefer not to assume such residual risk and obtain full insurance (*caveat venditor* contracts) but at a higher per unit price. Thus, if an unconstrained equilibrium in the product-cum-warranty market exists, it entails a separation of consumers into different risk classes. The intuition is that because the indifference curves in price-warranty space of the careful and careless consumers have different slopes, any putative single contract equilibrium will be broken by contracts which attract only careful customers. These separating contracts will contain *exculpatory clauses* which modify, explicitly or implicitly, either the scope of the warranty or limit the remedies for breach of existing warranties. Hence, if there is an equilibrium with unrestricted contracts, it will not entail the "pooling" of the consumers but rather the use of exculpatory clauses to separate consumers.

However, the imposition of a *caveat venditor* rule, with a prohibition on exculpatory clauses, precludes separation through incomplete warranties. Therefore, one consequence of a regulation imposing full producer liability is to force both careful and careless consumers to purchase the same contract; that is, they pay the same price for the product and both obtain full protection against product failure. This pooling contract makes careless consumers better off and careful consumers worse off as compared to their expected utilities in the separating "market" equilibrium. However, because the pooling contract has the risk-neutral firm bearing all the risk while in the unconstrained market equilibrium risk-averse consumers (albeit careful ones) bear some risk, one could argue that strict producer liability may lead to a "better" allocation of risk in the economy.

Thus far we have assumed that an equilibrium exists. An equilibrium will fail to exist if every separating set of contracts is Pareto dominated by a zero-profit pooling contract. In that case a pooling contract "breaks" the separating contract by attracting both careful and careless consumers. (A pooling contract is most likely to dominate a separating one if the proportion of careless consumers is small so that the "penalty" to a careful consumer from being grouped with the population as a whole is small relative to the cost, in terms of risk bearing, of the separating contract.) Although a Nash equilibrium does not exist, the sorting contracts are locally stable. That is, there exist sorting contracts which cannot be broken by other contracts in their neighborhood in price-warranty space. One might argue that this local stability is sufficient for these sorting contracts to arise in competitive markets where firms do not know the entire distribution of consumer types. In that case there are two possible market outcomes: a pooling contract and a set of sorting contracts. The former Pareto dominates the latter. A *caveat venditor* rule, with a prohibition on exculpatory clauses, by eliminating the possibility of separating contracts would (under these circumstances) be a Pareto improving change.

III. Conclusion

The usual debate over government restrictions on the prices or other terms of market transactions has revolved around equity versus efficiency arguments. We have confined our discussion to efficiency questions and shown that in markets with many sellers of a homogeneous product, but asymmetric information, government intervention may in some instances improve efficiency. Our analysis indicates that even in markets composed of a large number of agents, absolute freedom of contract does not inevitably conduce to an efficient allocation of resources. On the other hand, the presence of informational asymmetries does not provide a blanket justification for government restrictions on the terms of contracts.

REFERENCES

A. Braverman, "Consumer Search and Alternative Market Equilibria," *Rev. Econ. Stud.*, 1980, *47*, 487–502.

J. A. Ordover, "Products Liability in Markets with Heterogeneous Consumers," *J. Legal Stud.*, June 1979, *8*, 505–25.

R. W. Rosenthal, "A Model in which an Increase in the Number of Sellers Leads to a Higher Price," *Econometrica*, forthcoming.

M. Rothschild and J. E. Stiglitz, "Equilibrium in Competitive Insurance Markets," *Quart. J. Econ.*, Nov. 1976, *90*, 629–50.

J. E. Stiglitz and A. Weiss, "Credit Rationing in Markets with Imperfect Information, Part 1," *Amer. Econ. Rev.*, June 1981, forthcoming.

______ **and** ______, "Constraints as Incentive Devices, A Theory of Contingency Contracts: An Application to the Credit Market," mimeo. 1980.

H. Varian, "A Model of Sales," *Amer. Econ. Rev.*, Sept. 1980, *70*, 651–59.

The Economics of Privacy

By RICHARD A. POSNER*

The concept of "privacy" has received a good deal of attention from lawyers, political scientists, sociologists, philosophers and psychologists, but until recently very little from economists. This neglect is on the mend (see, for example, my 1978, 1979a articles and forthcoming book, chs. 9–11; George Stigler), and in this paper I will report on the economic research on privacy in which I and others have been engaged.

Some definitional clarification is necessary at the outset. Privacy is used today in at least three senses. First, it is used to mean the concealment of information; indeed, this is its most common meaning today. Second, it is used to mean peace and quiet, as when someone complains that telephone solicitations are an invasion of his privacy. Third, it is used as a synonym for freedom and autonomy; it is in this sense that the Supreme Court has used the word in subsuming the right to have an abortion under the right of privacy (see my 1979b article, pp. 190–200).

The third meaning of privacy need detain us only briefly. To affix the term privacy to human freedom and autonomy (as in Jack Hirshleifer) is simply to relabel an old subject—not to identify a new area for economic research. The second meaning of the word privacy set out above invites a slightly novel application of economics. It suggests an economic reason why certain (cerebral) workers have private offices and other (manual) workers do not, why aversion to noise is associated with rising education, and why certain low-level invasions of a person's "private space" (for example, shoving a person roughly but without hurting him) are tortious (see my forthcoming book, ch. 10). But the range of economic applications in this area seems limited.

The first meaning of privacy set out above—privacy as concealment of information—seems the most interesting from an economic standpoint. There is a rich and growing literature on the economics of information. It would seem that the same economic factors that determine search behavior by workers and consumers might also determine investments in obtaining, and in shielding, private information. This insight (emphasized in my 1978 article) provides the starting point for the economic analysis of privacy.

To relate the economics of privacy to the economics of information in as clear a fashion as possible, consider the example of the employer searching across employees and the employee searching across employers. The employer is looking for certain traits in an employee that may not be obvious, things like honesty, diligence, loyalty, and good physical and mental health. To the extent that the employee is deficient in one or more of these characteristics, he has an incentive—strictly analogous to the incentive of a seller of goods to conceal product defects—to conceal these deficiencies. That is, he has an incentive to invoke a "right of privacy" if the employer tries to "pry" into his private life.

The concealment of personal characteristics in the employment contest retards rather than promotes the efficient sorting of employees to employers. By reducing the amount of information available to the "buyer" in the labor market (the employer), it reduces the efficiency of that market. The analysis can easily be generalized, moreover, to other markets, some of them "noneconomic," in which private information is concealed. An example is the marriage "market." The efficient sorting of females to males in that market is impeded if either spouse conceals material personal information. The extended courtship that remains typical of the marriage market may be due in part to the efforts of prospective spouses to conceal their deficiencies from each other.

*University of Chicago Law School.

The length of the courtship is a social cost of concealment in the same way that additional investment in search by buyers is a social cost of fraud by sellers of goods.

The idea that fraud in "selling" oneself is just like fraud in the sale of goods is resisted on various grounds. It is sometimes argued that people will misuse private information—will attach excessive weight to knowledge that a prospective employee has a criminal record, or is a homosexual, or has a history of mental illness. However, the literature on the economics of nonmarket behavior suggests that people are rational even in nonmarket transactions such as marriage, and, in market transactions, even in regard to such apparently emotional factors as race and sex (see, for example, Gary Becker and Edmund Phelps). Therefore, there seems to be no solid basis for questioning the competence of individuals to attach appropriate (which will often be slight) weight to private information—at least if "appropriate" is equated with "efficient."

Various other arguments are made against the view that concealment of personal information is a form of fraud. It has been argued by Steven Shavell that such concealment provides a form of social insurance by buffering the wealth consequences of ill health, social misconduct, and other things that reduce wealth, since concealment may prevent the full wealth consequences of his condition or history from being visited on the individual. But concealment of adverse personal characteristics is surely an inefficient method of insurance; rather than spread costs widely, it shifts them from one small group to another. To take an extreme example, suppose that a teacher is allowed to conceal a history of sexual assaults on schoolchildren. The costs of concealment-as-insurance in this instance will not be spread throughout a large group but will instead be concentrated on the schoolchildren who become victims of this teacher in the future as a result of their (and the school board's) ignorance of his propensities.

It is also argued that disclosure of personal misconduct throws out of whack a carefully calibrated system of criminal sanctions; it increases the punishment for the crime, and reduces the prospects for rehabilitation of the criminal. But to foster concealment of a criminal past is to reduce the efficiency of the market for ex-criminals. It is more efficient to reduce sentences, or encourage rehabilitation by cash payments to the successfully rehabilitated criminal, than to force those who deal with the ex-criminal to do so in darkness.

More troubling to me is the argument (in Frank Easterbrook) from information overload. It is costly to assimilate heavy doses of information, much of it concerning facts of only peripheral relevance in deciding whether to hire or otherwise transact with an individual. This argument seems to me decisive against any rule requiring full disclosure of adverse personal information—on the model of the securities laws or the Truth-in-Lending Act. But it does not argue for granting legal protection to private facts about a person. And it is unlikely that the failure to create such rights just leads people to expend real resources on maintaining the secrecy of facts about themselves. No doubt such expenditures would be lower if there were such legal protection—but the same argument could be made on behalf of a proposal to give sellers a legally protected right to conceal adverse information about their product, and it is as unconvincing in the personal as it would be in the commercial context.

The arguments for privacy that I have reviewed are not absurd arguments. But, as just suggested, the same arguments could be made with equal force by a seller asking for the right to conceal defects in his product, yet would be accorded scant consideration in that context. The basic point I wish to assert is the symmetry between "selling" oneself and selling a product. If fraud is bad in the latter context (see Michael Darby and Edi Karni)—at least to the extent that we would not think it efficient to allow sellers to invoke the law's assistance in concealing defects in their goods—it is bad in the former context, and for the same reasons: it reduces the amount of information in the market, and hence the efficiency with which the market—whether the market for labor, or spouses, or friends—allocates resources.

TABLE 1—STATE PRIVACY STATUTE REGRESSIONS[a,b]

Constant	*TAX*	*PROG*	*RATIO1*	*RATIO2*	*TRAN*	*LTRAN*	*INC*	*LINC*	*MINO*	*MIG*	R^2
−1.013	−.0002						.0003		.0240	−.019	.12
(−.940)	(−.178)						(1.323)		(2.073)	(−.864)	
−1.428	−.0007	.030					.0004		.026	−.015	.14
(−1.214)	(−.594)	(.889)					(1.550)		(2.210)	(−.627)	
.060		.043	−8.292				.0003		.026	−.003	.18
(.051)		(1.322)	(−1.554)				(1.854)		(2.199)	(−.105)	
−1.033		.020		.327			.0003		.026	−.020	.13
(−1.010)		(.567)		(.034)			(1.601)		(2.201)	(−.913)	
−4.731					.023		.0005		.025	−.015	.23
(−2.677)					(2.491)		(2.992)		(2.284)	(−.717)	
−44.414						2.539		3.797	.026	−.005	.28
(−3.626)						(2.936)		(3.477)	(2.450)	(−.236)	

Source: My forthcoming book, ch. 10 (tab. 3).
Notes: *t*-statistics are shown in parentheses;

[a] Dependent variable = number of relevant categories in which state has enacted privacy statute.

[b] *Definitions of independent variables*: *TAX* = state taxes per capita, 1976; *PROG* = maximum state income tax rate minus minimum state income tax rate; *RATIO1* = ratio of per capita state and local expenditures excluding highway expenditures to state per capita income; *RATIO2* = *TAX*/*INC*; *TRAN* = ratio of total transfer payments to *INC*, 1976; *LTRAN* = natural logarithm of *TRAN*; *INC* = state per capita income, 1976; *LINC* = natural logarithm of *INC*; *MINO* = percentage black and Hispanic Americans in state; *MIG* = percentage of new residents since 1965.

My argument to this point will have seemed normative, but that is not its purpose. Once privacy is seen to reduce the efficiency of the marketplace, we are in a position to predict the effect of the recent wave of statutes, federal and state, protecting privacy, as by placing arrest records beyond a prospective employer's reach and credit histories beyond a prospective creditor's reach (see my 1979a article, pp. 41–50). If the analysis in this paper is correct, such statutes reduce wages and employment and increase interest rates.

The analysis in this paper is also suggestive with regard to the possible sources of privacy legislation. The principal beneficiaries of such legislation are people with more arrests or convictions, or poorer credit records (more judgments, bankruptcies, etc.), than the average person. These groups are presumably not cohesive enough to overcome the free-rider problems that plague efforts to form effective political coalitions, but they overlap strongly with racial and ethnic groups, namely black and Hispanic Americans, which are politically organized. Given laws that forbid discrimination against members of these racial and ethnic groups, it may be in their interest to press for passage of laws that also forbid "discrimination" against people with poor credit records and lengthy criminal records. If employers and creditors are unable to use these criteria to sift out poor employment risks and poor credit risks, respectively, a redistribution of wealth from whites to members of these racial and ethnic groups may result.

Table 1 presents some results broadly consistent with this theory. The dependent variable in the regressions reported there takes a value of 0 if the state has no privacy statute related to arrest, creditor or employment history, 1 if it has a statute in one of the categories, 2 if in two, etc. The key independent variable, *MINO*, measures the percentage black or Hispanic in the state. I add a variable which measures the amount of recent migration into the state (*MIG*) as a proxy for the social cost of privacy legislation, since the more often people change their residence the more difficult it is to obtain information about them that is useful in deciding whether to transact with them. Accordingly, the sign of *MIG* is expected to be negative. I also include a variable measuring per capita income in the state (*INC*) as a way of testing George Stigler's compassion theory of privacy legislation. Finally, since a state's resistance to redistributive legislation (as I regard privacy legislation) may be a (presumably negative) function of the amount of redistribution it already engages in, I include several variables that measure the state's other redistributive activities, such as per capita tax burden (*TAX*) and progressivity of the state income tax (*PROG*).

MINO is positive and significant in all of the regressions. *INC* is positive in all of the regressions too, as predicted by the compassion theory, but significant in only two. *MIG* has the right sign (negative), but is never significant. The variables measuring the amount of redistributive activity in the state are mostly insignificant and in one case have the wrong sign.

If I am correct that privacy legislation is redistributive and reduces rather than increases efficiency, it may seem puzzling, in light of recent economic literature claiming that the common law is efficient (see, for example, my 1977 book, part II), that the common law of torts recognizes and protects a "right to privacy." But on examination this right of privacy turns out to be consistent with the economic analysis in this paper (see my 1978 article, pp. 409–21). The tort right of privacy has four aspects. First, it prevents the use of a person's name or picture in advertising without his consent. The effect is to give a person a property right in his name and picture for purposes of advertising only (one cannot prevent a newspaper from publishing an unflattering picture of oneself in its news sections), and this maximizes the value of the name and picture in advertising without facilitating the use of the name or picture to mislead others.

Second, the tort law gives a person the right to prevent facts about him from being portrayed in a "false light." This right increases the amount of information in the market place. Third, the tort law prevents the obtaining of personal information by intrusive means, as by interfering with one's movements (an invasion of privacy in the second, and uncontroversial, sense discussed at the outset of this paper), or by eavesdropping. The economic objection to eavesdropping is that its principal effect is not to obtain information—not in the long run at least—but to reduce the effectiveness of communications. Knowing that people are overhearing my conversations, I will speak less frankly. The costs of communicating will be higher. Anyone familiar with the practical consequences of allowing student observers in faculty meetings will confirm the truth of this observation.

The only problematic aspect of the tort right of privacy is the right to prevent the publicizing of certain intimate facts about oneself. At first glance this right seems to be inconsistent with the economic analysis in this paper. Why would someone want to conceal a fact, except to mislead others in transacting with him? Examination of the cases shows, however, that the right is upheld in very few cases. Only in California do the courts allow a criminal record to be suppressed in a suit by the ex-criminal against the media. Elsewhere suppression is allowed only where the facts publicized have no possible value to potential transacting partners of the individual bringing the suit. Admittedly, *why* people should want to suppress such facts is mysterious from an economic standpoint.

To summarize, given the rash of recent privacy legislation and the high level of public as well as scholarly concern with privacy, the extension of the economic study of information to the privacy of information seems overdue. This paper and the work it reports on are far from definitive. But they suggest that here as in other areas of nonmarket behavior the economist has a distinctive and valuable contribution to make to social science scholarship.

REFERENCES

Gary S. Becker, *The Economics of Discrimination*, 2d. ed., Chicago; London 1971.

M. R. Darby and E. Karni, "Free Competition and the Optimal Amount of Fraud," *J. Law Econ.*, Apr. 1973, *16*, 67–88.

F. H. Easterbrook, "FOIA and the Value of Private Information," *J. Legal Stud.*, Dec. 1980, *9*, 775–800.

J. Hirshleifer, "Privacy: Its Origin, Function, and Future," *J. Legal Stud.*, Dec. 1980, *9*, 649–66.

E. Phelps, "The Statistical Theory of Racism and Sexism," *Amer. Econ. Rev.*, Sept. 1972, *62*, 659–61.

Richard A. Posner, *Economic Analysis of Law*, 2d ed., Chicago; Toronto 1977.

______, "The Right of Privacy," *Georgia Law Rev.*, Spring 1978, *12*, 393–422.

______, (1979a) "Privacy, Secrecy, and Reputation," *Buffalo Law Rev.*, 1979, *28*, 1–55.

______, (1979b) "The Uncertain Protection of Privacy by the Supreme Court," *Supreme Court Rev.*, 1979, 173–216.

______, *The Economics of Justice*, Cambridge, Mass. forthcoming.

G. J. Stigler, "An Introduction to Privacy in Economics and Politics," *J. Legal Stud.*, Dec 1980, *9*, 623–44.

Information Remedies for Consumer Protection

By HOWARD BEALES, RICHARD CRASWELL, AND STEVEN SALOP*

Consumer protection regulation has come under increasing fire from the Congress, courts, and the business community. In response, regulators have begun to innovate with market interventions that are more compatible with economic incentives. These incentive-compatible techniques include establishing property rights, mandating performance standards (instead of design standards), increasing competition, and encouraging and mandating information disclosure.

Information disclosure allows consumer self-protection, compatible with individual preferences. Information is also compatible with sellers' incentives, inducing them to compete on the basis of information disclosed. In addition, this competition increases the incentive to generate and disseminate additional product information, thereby repeating the cycle. In this way, information remedies rely on private economic incentives to achieve regulatory goals, rather than on expensive direct enforcement by the regulator.

Diagnosis of an information problem and evaluation of alternative remedies requires a number of steps: analysis of information production and distribution, identification of market failures and their implications for resource allocation in the information and product markets, and analysis of alternative remedies in light of these market failures.

I. Information Markets and Market Failures

The information market is diverse. Consumers produce prepurchase information themselves from direct inspection of commodity attributes. These attributes are desired for both their value in consumption and for their utility as signals of other valued attributes. Information recalled from memory and learned from experience is also useful, and essential for constructing signals. Experience may also be used to define conditions of contingency payments after further information is learned, as with warranties or trial periods. Consumers purchase information and certifications from a variety of intermediaries like newspapers and termite inspectors, and are given information by interested sellers.

The richness and competitiveness of information markets might suggest that it is not efficient to mandate dissemination of currently undisclosed information. However, market failures often prevent an efficient quantity and quality of product information from being provided. First, purchases by informed consumers generate a *market-perfecting* external benefit to uninformed consumers. Additional information induces sellers to compete for the patronage of informed consumers by offering better values. This induced competition also benefits those uninformed consumers who purchase randomly. Although perfect markets do not require all consumers to be perfectly informed, this externality implies that too little product information will generally be produced, even in a well-functioning information market.

There are other reasons to expect information markets to function imperfectly. First, information generation and dissemination has both *natural monopoly* problems (once generated, information can be disseminated at low marginal cost) and *free-rider* problems (buyers can resell purchased information to others). Second, firms with product market power (perhaps due to imperfect consumer information) may have

*Federal Trade Commission. The views expressed here are our own and do not necessarily reflect those of the Commission, individual Commissioners, or other staff. This is a brief report on ongoing research at FTC and elsewhere, and thus reflects the contributions of many of our colleagues. We acknowledge especially those who shared responsibility for the FTC's *Consumer Information Remedies* (1979) and *Post-Purchase Remedies* (1980) reports which examine these issues in more detail. References are cited in those volumes.

an incentive to act as *noisy monopolists* by exploiting or even creating uncertainty or false information.

II. Implied Product Market Failures

Information problems create imperfections in product markets and induce a variety of transactions costs and institutions to economize on them. First, if consumers are imperfectly informed, even small sellers can achieve a degree of *informational market power*, leading to monopolistic rather than perfect competition. For example, because the bereaved cannot easily shop among funeral homes, the industry is fragmented (each seller averages only 100 funerals per year) and prices are high. Spurious product differentiation and reputation premiums may raise prices for functionally equivalent brands. Finally, adverse selection and moral hazard can destroy markets altogether or lead to a low-price, low-quality "lemons" equilibrium. This may be a particular problem for warranties, where imperfect information, coupled with adverse selection and moral hazard, may lead to imperfect risk sharing and risk prevention.

Finally, the market responds by channeling competition towards more easily observable attributes and signals. If price is more easily observed than quality, competition will be skewed towards inexpensive, low-quality items. If experience suggests that a used car's exterior condition is a good signal for mechanical condition, "cleaner" cars will sell at a premium. As a result, sellers will be induced to overinvest in exterior condition to exploit the signal, possibly even destroying its predictive value in the process.

III. Information Remedies

Given a market failure, a number of alternative market intervention strategies may be designed. Remedies may improve the flow of truthful information to eliminate the cause of the problem, or they may act to offset the effects of the problems on the relevant product and information markets. Information strategies tend to be more compatible with incentives, less rigid, and do not require regulators to compromise diverse consumer preferences to a single standard.

Compatibility with sellers' incentives increases the likelihood of success. This is essential since the major benefits of a information program come from the market's indirect response, as firms compete for informed consumers. If the program is not a useful sales tool, then the market is less likely to respond. The need for effective communication is perhaps obvious, but its subtleties are often overlooked in practice; uncomprehended information that is ignored produces no benefits.

It is important to stress that information is inherently incomplete. Every statement can benefit from further elaboration or qualification. Thus, all information necessarily has the tendency or capacity to deceive. Government intervention must be limited to those that entail significant consumer injury and can be efficiently remedied without creating distortions or significant adverse side effects.

A. *Removing Information Restraints*

Private and governmental restrictions often tend to inhibit competition. For example, restrictions on advertising of professional services have raised prices. A diagnostician who refuses to make available diagnostic information may compel the consumer to purchase necessary treatments from him. Providing consumer access to such information enhances competition in the provision of treatments. Similarly, restrictions have been imposed by *Consumer Reports* to prevent retailers and manufacturers from using its ratings. Trademarks that have taken on generic meanings may also restrain the flow of information.

B. *Ensuring Truthful Information*

The FTC prohibits false claims and requires that firms have substantiation for advertised claims. While false information has negative economic value, requiring costly substantiation of truthful objective claims may give firms an incentive to substitute less valuable, unverifiable, subjective claims to escape the costs.

C. *Ensuring Complete Information*

Two types of incomplete information may be distinguished. Virtually all claims are incomplete in that they do not describe the other options available in the marketplace. For example, a firm may claim that its margarine has no cholesterol, without revealing that no margarine contains cholesterol. This converts a public good into a private one, and thereby gives an incentive to provide the information. On the other hand, omitting information about significant attributes (for example, cancer risks of cigarettes) may lead consumers to overestimate the value of a particular brand. Two common solutions to incomplete information are establishing a metric and requiring disclosure.

1. *Establishing Metrics*

A metric is a system for measuring the quantity of one or more product attributes across brands. The metric may be dichotomous, as with a definition (for example, "Walnut" means solid walnut, as opposed to veneer), or it may be continuous. Metrics reduce the cost of communicating by providing a uniform, easily comprehensible measurement. Thus, competing brands are more easily compared. The direct cost of imposing a metric include the one-time cost of establishing the index and the ongoing cost of testing the products to determine their scores. Testing costs are likely to increase with measurement precision.

Most metrics measure only a few product attributes. By easing communication about these attributes, the metric may increase the market's emphasis on them, at the expense of others. Particularly where unmeasured attributes are related to the measured one, either through production technology or preferences, increased emphasis on a newly observable attribute may lead to inefficient reductions in other attributes. The metric may become a signal for other unmeasured attributes. If the signal is imperfect, consumers may be misled if they rely solely on it.

Because sellers have an incentive to exploit the signal, it may become inappropriate over time. For example, the standard metric for nutrient composition of foods is the recommended dietary allowance (*RDA*). Because the role of many nutrients is incompletely u derstood and testing is expensive, some nutrients have no *RDA*. Instead, it was assumed that by obtaining the *RDA* of major nutrients from natural sources, sufficient amounts of trace elements would also be obtained. However, because manufacturers respond by fortifying natural products with synthetic vitamins, the assumed relationship between major and trace nutrients may no longer hold.

One solution to this problem is to measure more attributes. However, there is inevitably a tradeoff between the extensiveness of the measurements and their comprehensibility to consumers. Comprehensibility can sometimes be preserved by combining measures of different attributes into an overall summary index measure and perhaps collapsing the index into several discrete classes. An efficient index weights attributes in accordance with both consumer utilities and the precision of measurement. The usual problems of index numbers are always present when consumer preferences differ. In addition, strategic responses by producers may reduce the value of the index as discussed earlier. Collapsing an index into discrete classes may remove any incentive for marginal product improvements; a product that qualifies for the "best" class has no incentive for further improvements, if only the rating is observable.

Imposition of a single metric necessarily requires the exclusion of others. It is sometimes sufficient merely to establish the metric and make its use optional. Advantaged firms will voluntarily choose to disclose the metric, if it is an effective communication device. However, if the benefits of the metric depend on having the value for all products readily available, and if it is quite costly for each firm to test all products, it may be appropriate for the government to test or require that each firm test its own product and publish the results.

2. *Required Disclosure*

Disclosures may be *triggered* whenever a particular claim is made (for example, a claim about gas mileage triggers a requirement to disclose the *EPA* mileage estimate), or they may be *across the board* (for example, all cigarette ads must include a health warning). The need for disclosure requirements depends on the completeness of the total information environment and sellers' incentive to voluntarily disclose. If information is readily available elsewhere, then required disclosure is unnecessary. Requirements may sometimes be appropriate when a new metric is introduced to speed consumers' understanding of the new concept (for example, *R*-value), though such disclosures can often be terminated in a relative short time, once enough consumers learn the concept.

In contrast to a metric, disclosures tend to increase the cost of communication. Disclosures represent a tax on advertising which is collected in kind. The 1980-81 average cost of a prime time TV spot is approximately $3,000 per second of disclosure.

Triggered disclosures change the relative cost of different claims. If alternative claims are good substitutes as selling tools, claims which trigger the disclosure may be reduced. This may be inefficient if the claim is useful even absent the qualifying disclosure. Clearly, the actual magnitude of shifts in claims in response to disclosure requirements is an empirical question, and one which deserves further study. In contrast, when a disclosure applies to all advertising, substitution is impossible. However, the increase in the cost of advertising message may reduce the total amount of advertising.

Effective communication is essential for disclosures. For new information, consumers are likely to need a frame of reference to evaluate the information. The message must be consonant with the information processing capabilities of the target audience, and must consider the limitations of the medium in which it will be placed. As an alternative to actually writing the disclosure, a *performance standard* specifying a level of consumer awareness to be achieved gives firms an incentive to design the most cost-effective disclosure; this relieves the government of the task of writing effective advertising copy.

3. *Prohibiting Information*

Extensive triggered disclosures may amount to a virtual prohibition of the triggering claim, and thus redirect competition in permissible directions, as with professional advertising bans. Similarly, product differentiation competition may be reduced by a prohibition on certain types of product claims, though consumer preferences may be compromised by such a policy.

IV. Alternatives to Information Remedies

Information remedies are most likely to be the most effective solution to information problems. They deal with the cause of the problem, rather than its symptoms, and leave the market maximum flexibility. However, policymakers often consider remedies that act on the effect of imperfect information, such as altering contract terms or regulating products and prices. A mandated full warranty eliminates the need for information, since liability is shifted to the seller. A mandated cooling-off or trial period allows additional information to be gathered after purchase, but before a final commitment is made.

Clauses that void the remainder of the contract or are otherwise worthless to any informed purchaser seem ripe for prohibition. For example, the "Baldwin Piano" clause required that the product be returned to the factory at the consumer's expense to obtain warranty service.

Regulating products or prices can eliminate the need for information by requiring uniformity. If there are no choices, then there is little need for information about the options. Of course, if consumer preferences differ over the relevant attribute, a serious tradeoff must be balanced. Similarily, uniform price regulations eliminate competition. However, price regulations may encourage the flow of information by shifting competition from price to information services, as with resale price maintenance.

A Note on Efficiency vs. Distributional Equity in Legal Rulemaking: Should Distributional Equity Matter Given Optimal Income Taxation?

By STEVEN SHAVELL*

The question addressed in this paper is whether the choice of legal rules ought to be influenced by consideration of their redistributive effects. The answer to this question would, of course, be simple were it assumed that there was no difficulty in redistributing income, for then any socially undesirable distributional effect following from adoption of a particular rule could be undone by use of an appropriate redistributive tax scheme. Thus it would be best to choose legal rules only on the basis of criteria other than distributional equity, and, therefore, in the model to be studied here, to choose rules only on the basis of "efficiency."[1]

However, there is acknowledged difficulty in redistributing income; and such difficulty will be assumed below to be due solely to the adverse effect of an income tax on the incentive to work. In view of this problem, an otherwise socially optimal distribution of income generally would not be achievable, so that it might be expected that distributional equity as well as efficiency ought to enter into the choice of a legal rule. Suppose, for example, that the social preference is for income equality, but that, because an income-equalizing tax would severely depress work effort and lower the aggregate product to be shared, it would turn out to be best to employ only a mildly redistributive income tax. Consequently, one might suspect that it would be desirable to accomplish some further redistribution by giving an advantage under the law to those with relatively low income.

But this line of thought does not recognize that an attempt at redistribution through the choice over legal rules would involve the same sort of problem as exists under the income tax: If low-income individuals are treated relatively favorably in a legal setting, then there would be created a disincentive to work analogous to that associated with, say, a generous guaranteed minimum income under the tax schedule.

This suggests the result to be shown here, that despite imperfect ability to redistribute income through taxation, everyone would strictly prefer that legal rules be chosen only on the basis of their efficiency. After proving this result, I will comment on its interpretation and on its relationship to results in the literature on optimal income taxation.

I. The Model

Risk-neutral individuals are assumed to expend effort at work and to take care to prevent accident losses.[2] The individuals differ in their ability to earn income from work effort but not in their capacity to reduce accidents by taking care. Specifically, let w = work effort, $f(w)$ = disutility of work effort, a = ability, $p(a)$ = density of the population with ability a, y = income, $m(y)$ = density of the population earning income y, c = care taken to prevent accidents, $g(c)$ = disutility of care, $n(c)$ = density of the popu-

*Assistant professor, Harvard Law School. I wish to thank the National Science Foundation, grant no. SOC-76-20862, for research support, and Olivier Blanchard and Charles Fried for comments.

[1] It will be clear, though, that in an expanded version of the model of this paper, not only efficiency (the promotion of aggregate product or, equivalently, the reduction of aggregate losses plus prevention costs) but also other considerations that are not in strict logic income redistributional (for example, protection of personal rights and liberties) would generally affect the choice of legal rules.

[2] The terms "care" and "accident losses" are used only for concreteness; care could be more generally interpreted as any action having disutility, and accident losses as any outcome with a probability distribution affected by such an action.

lation taking care c, and $l(c;n)=$expected losses suffered by an individual given c and n.

For simplicity, work effort and income are assumed to be linearly related,

$$y=aw \tag{1}$$

The government sets a tax based on income; because the government is presumed to be unable to observe directly either ability or work effort, the tax cannot be based on either of those variables. Let $t(y)=$income tax given y. Since the government is assumed to return in the aggregate all taxes collected, the tax schedule must satisfy[3]

$$\int t(y)p(a)da=0 \tag{2}$$

where y should be understood to be a function of a. This function and the density m will be determined below.

The care an individual takes is assumed to affect the probability distribution of losses suffered by others through its effect on the density function n, and it may affect as well the distribution of losses suffered by the individual himself. Losses are treated as a subtraction from income and therefore, since individuals are risk neutral, only expected losses l are considered.

A liability rule specifies how much an individual who is involved in an accident pays or receives in damages. Such a rule is allowed to depend on the levels of care of the involved parties, on their incomes, and on the magnitude of harm done. However, there will not be a need to consider liability rules explicitly. All that will matter below is the relationship between an individual's expected damage payments and his level of care, and this is implicitly determined by choice of a liability rule. Let $d(c,y;m,n)=$ expected net damages paid by an individual under a liability rule given c, y, m, and n.[4] (If a liability rule does not depend on income, then expected damages will be denoted by the simpler form $d(c;n)$.) Since under a liability rule, what one individual pays another receives, aggregate net liability payments must be zero,

$$\int d(c,y;m,n)p(a)da=0 \tag{3}$$

Here, c and y are to be understood as functions of a.

The expected position of an individual of ability a may now be written, given his work effort and level of care,

$$y-t(y)-l(c;n) \\ -d(c,y;m,n)-f(w)-g(c) \tag{4}$$

where $y=aw$. The first four terms comprise expected income. And, as will be explained subsequently, the assumption of the separable form of the last two terms, the disutilities of work effort and of care, is necessary to the result to be proved.

It is assumed that an individual chooses (the unique) levels of work effort and of care that maximize expected utility (4), while taking the tax schedule, the liability rule (and thus d), and the population distributions (m,n) as fixed. This defines w, c and y as functions of a; and from the functions c and y, induced population distributions of care and of income can be derived.[5] It should be observed from (4) that given m and n, an individual's choice of c is completely determined by his choice of y. Thus c may be written as $c(y;m,n)$ and (4) may be rewritten as

$$y-t(y)-l(c(y;m,n);n) \\ -d(c(y;m,n),y;m,n) \\ -f(w)-g(c(y;m,n)) \tag{5}$$

where $y=aw$.

[3] Note from (2) that some individuals will generally pay a negative tax, i.e., receive payments.

[4] A negative value of d corresponds to expected receipt of damages.

[5] For example, given care as a function of ability (and knowing the density p of ability), we can determine the density of care.

It is also assumed that an equilibrium exists for any tax schedule and liability rule: Given a t and a d, there exist m and n such that if these are taken as fixed, the induced distribution of income is in fact m, and that of care is in fact n.

Consider now the problem of minimizing expected accident losses plus the cost of care,

$$(6)\qquad \int [l(c;n)+g(c)]p(a)da$$

It can be shown under general conditions that this problem is solved by having all individuals exercise a level of care c^*, to be called the efficient level of care.[6] And it can then be shown that there exist liability rules depending at most on harm done and levels of care (and thus not on income) which induce parties to take the efficient level of care. Such liability rules will be called efficient.[7] Let n^* denote the distribution of care levels under an efficient liability rule, i.e., n^* denotes the degenerate distribution under which all individuals take care c^*.

II. Proof of the result

The result to be established is as follows: *Suppose that under a liability rule some* (a positive fraction) *or all individuals are led to exercise an inefficient level of care* (perhaps because the rule is to some extent based on income). *Then by adoption instead of an efficient liability rule and by appropriate modification of the income tax schedule, everyone can be made strictly better off.*

To prove the result, let us use a hat to denote variables and functions in the situation under the inefficient liability rule and an asterisk to denote variables and functions in a new situation—to be constructed—under an efficient liability rule. (Since in the new situation the liability rule is efficient, there is no conflict with the previous definitions of c^* and n^*.) Define the new tax schedule by

$$\begin{aligned}(7)\quad t^*(y)=\hat{t}(y)&-\big[l(c^*;n^*)\\ &-l(\hat{c}(y;\hat{m},\hat{n});\hat{n})\big]\\ &-\big[d(c^*;n^*)-d(\hat{c}(y;\hat{m},\hat{n}),y;\hat{m},\hat{n})\big]\\ &-\big[g(c^*)-g(\hat{c}(y;\hat{m},\hat{n}))\big]-s^*\end{aligned}$$

where

$$\begin{aligned}(8)\quad s^*=&\int\big[l(\hat{c};\hat{n})+g(\hat{c})\big]p(a)da\\ &-\big[l(c^*,n^*)+g(c^*)\big]\end{aligned}$$

(Note here that $\hat{c}$ is the function $\hat{c}(a)$ relating care to ability obtaining in the original situation.) Thus s^* is the expected savings in accident losses plus prevention costs to be had by use of an efficient liability rule. This savings is positive by assumption.

An individual's expected utility as a function of w and c under the efficient liability rule and the new tax is (see (4))

$$\begin{aligned}(9)\quad &y-t^*(y)-l(c;n^*)-d(c;n^*)-f(w)\\ &-g(c)=y-\hat{t}(y)-l(\hat{c}(y;\hat{m},\hat{n});\hat{n})\\ &-d(\hat{c}(y;\hat{m},\hat{n}),y;\hat{m},\hat{n})-f(w)\\ &-g(\hat{c}(y;\hat{m},\hat{n}))\\ &-\big[l(c;n^*)+d(c;n^*)+g(c)\\ &-l(c^*;n^*)-d(c^*;n^*)-g(c^*)\big]+s^*\end{aligned}$$

where $y=aw$. However, since all individuals will choose c^*, the term in brackets equals zero. Consequently, (9) reduces to

$$\begin{aligned}(10)\quad &\big[y-\hat{t}(y)-l(\hat{c}(y;\hat{m},\hat{n});\hat{n})\\ &-d(\hat{c}(y;\hat{m},\hat{n}),y;\hat{m},\hat{n})-\\ &f(w)-g(\hat{c}(y;\hat{m},\hat{n}))\big]+s^*\end{aligned}$$

And since the term in brackets is the ex-

[6] Moreover, it should be noted that minimization of (6) is a necessary condition for achieving a (first best) Pareto optimum.

[7] For example, in models of accidents like those in Peter Diamond and my earlier article, strict liability with a defense of contributory negligence or the negligence rule would be efficient.

pected utility function in the original situation (see (5)), and since $s^* > 0$, all individuals are strictly better off.

It remains to show that the government breaks even under the new tax schedule, i.e., (2) is satisfied by t^*. Now because (10) differs from the term in brackets by a constant, it follows that individuals choose the same levels of work effort as they did under $\hat{t}$. Accordingly, gross income is the same function of a, i.e., $y^*(a) = y(a)$, and we have (using also (7), (2), (8), and (3)),

$$(11)\quad \int t^*(y^*)p(a)da = \int t^*(\hat{y})p(a)da$$

$$= \int \hat{t}(\hat{y})p(a)da + \Bigg[\int [\, l(\hat{c}; \hat{n})$$

$$+ g(\hat{c})]\, p(a)da - l(c^*, n^*) - g(c^*) - s^* \Bigg]$$

$$- d(c^*; n^*) + \int d(\hat{c}(\hat{y}; \hat{m}, \hat{n}), \hat{y};$$

$$\hat{m}, \hat{n}) p(a)da = 0 + 0 - 0 + 0 = 0$$

III. Comments

(a) A familiar point of qualification about results such as the one proved here probably bears repeating, namely that if the income tax would not be altered on adoption of new liability rules, then in strict logic the argument given for use of efficient rules does not apply. Now, of course, no one would really expect the income tax structure to be adjusted in response to each and every change in legal rules (much less to individual changes in other domains), for this would be impractical. Therefore, one's attitude toward the result under discussion will depend on his expectation that the income tax would be (or could be) altered in response to changes in legal rules whenever these changes resulted in a "sufficiently important" shift in the distribution of income.

A second point of qualification concerns the possibility that income might be correlated with certain unobservable individual characteristics which ought to lead to favorable legal treatment. If so, income might be employed as a proxy for these characteristics and thereby justifiably influence legal outcomes. For example, suppose that poor individuals' decisions in the marketplace are not as well informed as those of individuals with moderate or high incomes. Then, to the extent that lack of consumer knowledge should influence legal outcomes but cannot be observed by the courts, income could be used as an indicator of lack of knowledge and thus could affect legal outcomes in a desirable way.

A similar point concerns the role of the payment of money damages as social insurance against loss caused by others. To the extent that this role of legal rules is important and that the poor have a greater need for insurance (because of decreasing absolute risk aversion), they might receive favorable legal treatment.

(b) The result shown here is closely related to two results in the literature on income taxation and its adverse effect on the incentive to work. The first result, in James Mirrlees, concerns the use of linear commodity taxation (a fixed tax per unit of the commodity purchased) given simultaneous use of optimal income taxation. Mirrlees shows (among other things) that if the demand for a commodity is independent of income, then it should not be taxed. In other words, there is no scope for beneficial redistribution through linear commodity taxation given optimal income taxation. This is clearly similar to what was proved here, for the motive to take care was independent of income.[8] The second result, in A. Hylland and Richard Zeckhauser, has to do with the choice among government projects, again given simultaneous use of optimal income taxation. Hylland and Zeckhauser consider a model in which there is one produced good; and a government project is

[8] Mirrlees also shows that if demand for a commodity is affected by income, then it may be desirable to impose a tax; when, for example, demand increases with income, a positive linear commodity tax would be used. By analogy, it would be expected that a similar result would hold in regard to legal rules (if the cost of taking care or if the type of accident differed in a systematic way with income).

identified with a function specifying the net amount (positive or negative) of the good to be enjoyed given income. They show that the project that ought to be adopted is the one with highest aggregate net benefits. This is similar to the result of this paper, for the choice of a legal rule may be likened to the choice of a project.[9]

[9] The principal difference between the problem analyzed here and that analyzed by Hylland and Zeckhauser (and by Mirrlees) is the externality associated with individuals' choice of care.

REFERENCES

P. A. Diamond, "Single Activity Accidents," *J. Legal Stud.*, Jan. 1974, *3*, 107–64.

A. Hylland and R. M. Zeckhauser, "Distributional Objectives Should Affect Taxes but not Program Choice or Design," *Scandinavian J. Econ.*, No. 2, 1979, *81*, 264–84.

J. A. Mirrlees, "Optimal Tax Theory: A Synthesis," *J. Public Econ.*, Nov. 1976, *6*, 327–58.

S. Shavell, "Strict Liability versus Negligence," *J. Legal Stud.*, Jan. 1980, *9*, 1–25.

INFLATION AND THE TAX SYSTEM

Inflation and the Tax Treatment of Firm Behavior

By ALAN J. AUERBACH*

In the past decade, economists have begun to realize that inflation, even when fully anticipated, constitutes a great deal more than a tax on money balances. The primary reason for inflation's wider impact is the existence of a tax system designed with stable prices in mind. This paper offers a brief summary of the effects of inflation on the tax treatment of the firm, focusing on four important decisions the firm makes: the scale of investment; the method of finance; the durability of assets used in production; and the holding period of these assets.

There are a number of interesting and related issues which cannot be covered in a paper of this length. As I will be considering inflation that is both uniform and fully anticipated, questions concerning the behavior of the firm in response to uncertainty about inflation, or to a concommitant change in relative prices, will not arise.

I. The Model

Let us consider a simple model of a corporation which uses a single type of capital good in producing one type of output. The firm seeks to maximize the wealth of its shareholders, who discount after-tax cash flows at rate e and are subject to personal taxes on dividends at rate θ, and capital gains, at an accrual-equivalent rate c. The firm pays taxes at rate τ on corporate profits, which are calculated by deducting interest payments and depreciation allowances from gross cash flows. The nominal interest rate is i, and b is the fraction of capital structure that the firm chooses to devote to debt.

All capital goods are assumed to have service patterns which decline exponentially; the rate of decay δ is indicative of how durable the asset is. The price, relative to that of output sold concurrently, of a unit of capital of type δ yielding a certain standard level of capital services is $q(\delta)$. All prices inflate at rate π. As the purpose of this paper is to focus on the specific impact of inflation, I shall consider the simple case in which depreciation allowances accorded assets would reflect actual economic depreciation in the absence of inflation, but which are based on historic cost. This implies that the nominal depreciation allowance received by an asset of age t and type δ is $\delta e^{-\delta t}$ times its original purchase price. I also omit the investment tax credit in the interest of simplicity.

Firms not only choose the durability of the assets used, but the length of time T that they are held before being sold and replaced. Upon such resale, firms are taxed at rate $\gamma \leqslant \tau$ on the difference between sale price and basis (the nominal value of remaining depreciation deductions).

As shown in the Appendix, the firm's optimal behavior may be viewed as a two-stage process. In the first stage, it chooses the decay rate δ, the holding period T, and the debt-value ratio b to minimize the "user cost" of capital, which is the shadow rental price of capital goods. In the second stage, the firm invests until the marginal product of capital goods equals this minimized cost. For given values of δ, T, and b, the user cost is

$$(1)\quad C=\frac{q(\delta)(\rho+\delta)}{(1-\tau)}\left[(1-\tau z)+(\gamma-\tau z)\times(1-e^{-\pi T})\left(\frac{e^{-(\rho+\delta)T}}{1-e^{-(\rho+\delta)T}}\right)\right]$$

where

$$(2)\quad \rho=\frac{bi(1-\tau)(1-\theta)+(1-b)e}{b(1-\theta)+(1-b)(1-c)}-\pi$$

*Assistant professor of economics, Harvard University, and Research Associate, National Bureau of Economic Research.

may be interpreted as the real after-tax cost of funds to the firm and

$$(3)\quad z=\int_0^{\infty} e^{-(\rho+\pi)t}\delta e^{-\delta t}dt=\frac{\delta}{\rho+\pi+\delta}$$

is the present value of depreciation allowances accruing to an initial investment of one dollar which is never resold (discounted at the nominal discount rate $\rho+\pi$ because allowances are in nominal terms). To get an intuitive sense of what ρ represents, note that when $b=1$, $\rho=i(1-\tau)-\pi$, the interest rate net of tax deductions and inflation; when $b=0$, $\rho=(e/1-c)-\pi$.

Equation (1) differs from the standard formula for user cost because it explicitly accounts for the tax treatment of the disposal of assets by resale. It reduces to the basic formula when $T=\infty$.

II. The Effects of Inflation

A. *Asset Holding Period*

In a more general model than that considered here, firms might find it optimal to sell and replace assets of a certain vintage, rather than use them until fully exhausted, even in a world without taxes. In the current model, all assets are identical in productive characteristics, so such behavior could have no real consequences.

However, the introduction of taxes may cause assets identical in productive characteristics to differ in another sense. If depreciation allowances are accelerated, an asset declines in value faster than would be dictated by its decline in productivity alone. This is because it is now really two "assets": one that produces capital services, and one that "produces" depreciation deductions, the second declining in value more rapidly than the first. However, if the asset is sold, under current *U.S.* law the depreciation allowances that remain are not transferred. Rather, the sale price is used as a new basis for depreciation deductions. Thus, if the depreciation schedule is accelerated, the asset transfer will increase the value of remaining deductions and generate an increase in the value of the asset. This is countered by the fact that the seller must pay a tax at rate γ on the difference between sale price and basis (the nominal value of remaining depreciation allowances). The rate γ simply equals τ for equipment, but for structures is actually a weighted average of the ordinary corporate rate and the lower corporate capital gains rate; the ordinary rate is applied only to the amount by which the asset's basis falls short of that which would have obtained had straight-line tax depreciation been used. (This practice is technically referred to as the "recapture" of "excess" depreciation, though such a designation is rather inappropriate.) Imagining a firm selling the asset to itself, we can see that it must weigh the increased value of depreciation allowances against the tax liability incurred on transfer.

When there is economic depreciation of assets, as I have assumed in this analysis, such a distortion disappears; basis and sale price would be identical and turnover would have no real impact on the firm. However, inflation once again introduces the same divergence caused by accelerated depreciation. Historic cost depreciation implies that turnover provides a step-up in basis, generating both an increase in the value of future depreciation deductions and an immediate tax liability.

This effect is represented by the second term in brackets in the cost of capital expression in equation (1). This term increases or decreases with T according to whether the turnover tax γ is less than or greater than the present value of tax deductions. Since, for structures, γ is approximately equal to τ, currently .46, for small values of T, and approximately equal to the corporate capital gains rate, currently .28, for T large (because the fraction of sale price less basis "recaptured" declines over time), the optimal holding period T, with positive inflation, will be zero, infinite, or somewhere in between according to whether $z\geqslant 1$, $z\leqslant .28/.46$ or $1>z>.28/.46$. The first condition is never met, and the second requires that $\delta<(\rho+\pi)x(.28/.18)$. For a nominal discount rate of .10, this critical value of δ is .156, much higher than the rate of depreciation for any general category of structures.

Since $\gamma \equiv \tau$ for equipment, an optimal holding period less than infinity never obtains. Thus, for most assets, inflation will encourage holding assets, despite their inflation-eroded depreciation allowances, rather than replacing them.

B. *Debt-Equity Ratio*

As the Modigliani-Miller theorem shows, the choice of debt-equity ratio is of no consequence in a taxless world under a variety of circumstances, and debt dominates equity with a corporate tax but no personal taxes. However, in reality, holders of debt and equity pay taxes, too, and because of the favorable tax treatment of capital gains, the personal tax rate on debt income is higher for any given individual than the tax on equity income. Thus, the choice between debt and equity depends on the relative magnitudes of the corporate tax rate, τ, the capital gains rate, c, and the personal tax rate, θ. As I discussed in an earlier paper (1979b), the debt-equity choice is knife-edged if all investors possess the same tax rates, even in the presence of short sale constraints on individuals. However, with progressive taxes, an interior solution is possible in which firms are indifferent between debt and equity and individuals are specialized in clienteles.

To examine the effect of inflation on the debt-equity decision, I rewrite equation (2) by replacing i and e with the real, after-tax returns to holders of equity and debt, $e_N = e - \pi$, and $i_N = i(1-\theta') - \pi$, where θ' is the personal tax rate of those who hold debt, and not necessarily equal to θ. Equation (2) becomes

$$(4) \quad \rho = [b(1-\theta)+(1-b)(1-c)]^{-1} \times \left\{ \left[b i_N (1-\tau)\left(\frac{1-\theta}{1-\theta'}\right) + (1-b)e_N \right] + \pi \left[b\left(\frac{1-\theta}{1-\theta'}\right)(\theta'-\tau) + (1-b)c \right] \right\}$$

For given underlying real rates of return e_N and i_N, inflation influences the real cost of funds ρ in three ways, depicted in the term multiplying π in (4). First, corporations can deduct at rate τ the inflation premium component of the nominal interest rate; second, bondholders must pay tax rate θ' on the same amount. Thus, for i_N given, debt becomes cheaper to the firm as inflation increases if $\tau > \theta'$, and more expensive if $\tau < \theta'$. Although τ is directly observable, θ' is not, because individual tax rates differ; estimates of θ' vary considerably. From a comparison of returns on tax-exempt and taxable long-term debt, Roger Gordon and Burton Malkiel estimate θ' to have been approximately 22.5 percent in 1978. Using flow of funds data to identify holders of debt and calculate θ' directly, Martin Feldstein and Lawrence Summers arrive at a value of 42 percent for 1977. It is thus unclear to what extent inflation reduces the effective tax rate on debt, if at all, though it seems likely that no appreciable additional tax burden is introduced.

The final influence of inflation on ρ is through the taxation of nominal rather than real capital gains. Here, there is no question about the direction of the effect; for e_N given, equity becomes more expensive. Estimates of c, like those of θ', are not very accurate, though c may very well be under 10 percent, as suggested by Martin Bailey. (Remember that c is the accrual-equivalent of the tax rate on realizations.) Thus, for given values of e_N and i_N, the likely effect of inflation is to make debt a cheaper source of finance, and equity more expensive, encouraging greater use of the former. Of course, the general equilibrium effect of inflation on b is more complicated, for it must also depend on the behavior of e_N and i_N.

C. *Choice of Asset Life*

Assuming the choice of asset durability to be among values of δ in the "normal" range where the optimal holding period T is infinite, the cost of capital for given b and δ may be written more simply as

$$(5) \quad C = \frac{q(\delta)(\rho+\delta)}{(1-\tau)}(1-\tau z) = q(\delta)\left[\frac{\rho}{1-\tau} + \delta + \frac{\tau\pi z}{1-\tau}\right]$$

Expression (5) shows that the user cost per dollar of capital consists of three terms: the gross of tax real firm discount rate, the rate of asset decay, and the rate of decline due to inflation in the value of the nominally denominated "asset" representing the present value of the stream of depreciation allowances.

Perhaps a commonly held belief is that this "inflation tax" on depreciation allowances weighs more heavily on longer-lived assets which have to wait longer to collect their depreciation allowances. This view is incorrect (see my 1979a article). For any given value of ρ, the required internal rate of return before taxes on an asset of type δ is

$$(6) \quad \nu(\delta)=\frac{C(\delta)}{q(\delta)}-\delta=(\rho+\tau\pi z)/(1-\tau)$$

It is evident that while inflation raises this rate for all values of δ, the rate of change increases monotonically with δ; the size of the inflation tax declines with asset durability.

It is important to realize that just as the increase in the tax burden on equity relative to debt does not *necessarily* imply that inflation will lead to increased leverage in a full general equilibrium model, the heavier rate of tax on short-lived assets needn't imply that a smaller value of δ will result from inflation. The ultimate answer depends on the behavior of the real after-tax return ρ. If ρ is fixed, there are two offsetting effects which determine the optimal δ. The relatively higher tax rate on short-lived assets will favor the choice of a small value of δ. However, the general increase in all tax rates, with the resulting higher before-tax rate of return, favors the choice of short-lived assets with large values of δ. As has been pointed out by Richard Kopcke, the total effect on the choice of δ is ambiguous, as can be seen from considering the effect of π on the cost of capital. On the other hand, if ρ decreases with the increase in inflation, as Patric Hendershott has suggested, this second effect favoring short-lived investment is lessened.

D. *Investment Scale*

The scale of investment depends on the cost of capital. If we hold constant the underlying rates of return to investors, e_N and i_N, and the other decision variables of the firm, b, δ and T, then the likely effect of inflation, as discussed by Feldstein and Summers, will be an increase in user cost and a drag on investment. The effect on ρ will be ambiguous but small relative to the increase in the inflation tax on depreciation allowances. For example, for representative values of the relevant parameters ($\theta=.4$, $\theta'=.3$, $\tau=.46$, $c=.1$, $T=\infty$, $b=.3$, $\delta=.1$, $\rho=.04$, and $\pi=.06$) an increase in the rate of inflation of $\Delta\pi$ raises ρ by $.036\Delta\pi$, while $\tau\pi z$ increases by $.161\Delta\pi$.

However two important qualifications are necessary. First, if the real after-tax rates of return e_N and i_N fall as a result of inflation, as some theory and evidence suggests, ρ will increase less (or decrease more) and so will user cost, than has been proposed. Moreover, to the extent that firms can alter their debt-equity ratio and choice of asset durability, this must also diminish the increase in user cost. The answer to how inflation affects the scale of investment thus depends in part on a number of empirical magnitudes about which more information should be acquired.

Appendix: The Firm's Optimizing Behavior

Let us assume the firm produces output with the concave production function $F(K)$, where K is the capital stock on hand. The firm seeks to maximize the wealth of its owners as represented by the present value of net cash flows, discounted at the equity rate e. As demonstrated in my earlier paper (1979b), this is equivalent to choosing b to minimize ρ (as presented in equation (2) in the text), and then maximizing the present value, calculated with discount rate $\rho+\pi$, of flows to the firm before interest payments and debt issues. Letting I_t be the physical investment in capital at time t, this present value is

$$\text{(A1)}\quad v=\int_0^{\infty} e^{-(\rho+\pi)t} \times \left.(1-\tau)e^{\pi t}F\left(\int_{-\infty}^{t} I_s e^{-\delta(t-s)}ds\right) - e^{\pi t}q(\delta)I_t(1-x)\right]dt$$

where x is the present value of depreciation allowances times τ plus turnover tax payments per dollar of investment. If each asset is turned over every T years, the present value of depreciation deductions it receives per initial dollar is

$$\text{(A2)}\quad \int_0^T e^{-(\rho+\pi)t}\delta e^{-\delta t}dt + e^{-\rho T}\int_T^{2T} e^{-(\rho+\pi)(t-T)}\delta e^{-\delta t}dt + \ldots = z\left(\frac{1-e^{-(\rho+\pi+\delta)T}}{1-e^{-(\rho+\delta)T}}\right)$$

which exceeds z because of the step-up in basis every T years. The present value of turnover tax payments is

$$\text{(A3)}\quad \gamma\left[e^{-\rho T}e^{-\delta T}(1-e^{-\pi T}) + e^{-2\rho T}e^{-2\delta T}(1-e^{-\pi T}) + \ldots\right] = \gamma\left(\frac{e^{-(\rho+\delta)T}}{1-e^{-(\rho+\delta)T}}\right)(1-e^{-\pi T})$$

Combining (A2) and (A3) yields

$$\text{(A4)}\quad x=\tau z+(\tau z-\gamma) \times (1-e^{-\pi T})\left(\frac{e^{-(\rho+\delta)T}}{1-e^{-(\rho+\delta)T}}\right)$$

Insertion of this value of x into (A1) and differentiating v with respect to I_t yields the requirement that the marginal product of capital F' equals C, as represented in equation (1) in the text. Differentiation of v with respect to δ and T yields the conditions that $\partial c/\partial\delta$ and $\partial C/\partial T$ should equal zero.

REFERENCES

A. J. Auerbach, "Inflation and the Choice of Asset Life," *J. Polit. Econ.*, June 1979, *87*, 621–38.

______, "Wealth Maximization and the Cost of Capital," *Quart. J. Econ*, Aug. 1979, *93*, 434–46.

M. J. Bailey, "Capital gains and Income Taxation," in Arnold C. Harberger and Martin J. Bailey, eds., *The Taxation of Income from Capital*, Washington 1969, 11–49.

M. Feldstein and L. Summers, "Inflation and the Taxation of Capital Income in the Corporate Sector," *Nat. Tax. J.*, Dec. 1979, *32*, 445–470.

R. H. Gordon and B. G. Malkiel, "Taxation and Corporation Finance," memo. no. 31, Princeton Univ. Financial Research Center 1980.

P. H. Hendershott, "The Decline in Aggregate Share Values: Inflation, Taxation, Risk and Profitability," paper presented at Nat. Bur. Econ. Res. Conference on Taxation, Nov. 1979.

R. W. Kopcke, "Inflation, Corporate Income Taxation, and the Demand for Capital Assets," *J. Polit. Econ.*, forthcoming.

Private Pensions and Inflation

By MARTIN FELDSTEIN*

Much of the recent discussion about the relation between pensions and inflation has emphasized the adverse impact that the unexpected rise in inflation has had on pension recipients and on the performance of pension funds. In contrast, the present paper focuses on the way that pensions are likely to evolve in response to the expectation of continued inflation in the future and to the uncertainty about the rate of inflation.

The unfortunate effects that occurred when inflation caught pensioners and pension fund managers by surprise should not be confused with an inability to adjust to future conditions, even uncertain future conditions. As I shall explain, the persistence of a high rate of inflation is likely to increase the share of total saving that goes into private pensions. Since the tax treatment of pension contributions allows individuals to save in this way for retirement on the same terms that they would under a consumption tax,[1] the existence of the private pension system may be one of a few things that prevents the national saving rate from going even lower in the current inflationary environment.

I. Expected Inflation, Asset Yields, and Pension Savings

This section analyzes a steady state in which expected asset yields are constant and expectations are fulfilled.

A. *An Equity-Only Economy*

Consider first an economy in which there is no debt. Corporations finance their investments by issuing equity and reinvesting retained earnings. Pension funds invest only in corporate equities.

With the existing *U.S.* tax rules, inflation would unambiguously reduce the return that firms could earn on equity.[2] Although this extra inflation-induced tax at the corporate level affects both households and pensions, the after-tax yield to households is further depressed by the tax that must be paid on nominal capital gains.[3] Thus inflation unambiguously reduces the yield on equities to pensions by less than the reduction in the yield on equities held directly by households.

The increase in the yield spread in favor of pensions should cause pensions to have a larger share of the total private saving. Of course, because the effective tax rate on capital gains is relatively small, the advantage of the switch is also small and the redistribution of assets would not be large. It would nevertheless be in the direction of increasing and strengthening the role of private pensions.

B. *A Debt-Only Economy*

Consider now a debt-only economy, or at least an economy in which all marginal corporate investments are financed by debt, and pension funds invest exclusively in debt. The effect of a change in inflation in such an economy is both more complex and more

*Professor of economics, Harvard University, and President, National Bureau of Economic Research. A more complete version of this paper is available as NBER Working Paper No. 568. It is part of the NBER Study of Private and Public Pensions. This work has benefited from discussion with other participants in that study, especially Zvi Bodie. I am grateful to the National Science Foundation and the NBER for support. The views expressed here are my own and should not be attributed to any organization.

[1] This is true only if the limits on pension contributions or pension benefits are not binding.

[2] This result form historic cost depreciation and the remaining use of FIFO inventory accounting. The mitigating effect of deducting nominal interest payments is temporarily ruled out by assumption.

[3] An explicit analysis of the effect of inflation on equity yields of households and pension funds is presented in my 1980a, b articles.

dramatic than in the all-equity economy and depends on the way that the interest rate responds to inflation.

The stylized "fact" that the real interest rate remains constant is a useful starting point.[4] Since pension funds are not taxed, a constant real rate of interest implies that they earn a constant real *net* rate of interest. In contrast, the real net yield earned by households drops sharply for each percentage point rise in the rate of inflation. If the rate of tax paid by households is θ, the real net interest rate is $r_h=(1-\theta)i-\pi$ where i is the nominal rate, π is the rate of inflation, and the h subscript refers to the return to households. Thus $dr_h/d\pi=(1-\theta)(di/d\pi)-1=-\theta$ when $di/d\pi=1$. If the individual tax rate averages 30 percent ($\theta=.3$), the real net interest rate would fall from 2.8 percent when there is a 4 percent nominal interest rate and no inflation to 1.0 percent when there is 6 percent inflation and a 10 percent interest rate. The yield difference between pension funds (earning a real 4 percent in both cases) and direct saving nearly triples from 1.2 percent to 3.0 percent.

An alternative simple view is that the nominal interest rate is determined as the yield that corporations can afford to pay with a fixed real return on capital (f').[5] The firm's first-order condition is readily shown to be[6]

$$f'+\pi=\tau f'+\tau\mu\pi+(1-\tau)i \tag{1}$$

The nominal pre-tax return per unit of capital ($f'+\pi$) is exhausted by the tax on the real return on capital ($\tau f'$ where τ is the corporate tax rate) plus the tax on the nominal gains that results from historic cost depreciation ($\tau\mu\pi$ where μ measures the increase in taxable profits on a unit of capital per percentage point increase in inflation),[7] plus the maximum net yield on debt ($(1-\tau)i$). Thus the nominal interest rate is given by

$$i=f'+\frac{1-\mu\tau}{1-\tau}\pi \tag{2}$$

With f' constant, $di/d\pi=(1-\mu\tau)/(1-\tau)$. The value of μ depends on the rate of inflation but a reasonable approximation is $\mu=0.5$ (see my 1980b article, especially the Appendix). Thus with the current 46 percent corporate tax rate, $di/d\pi=1.43$. The nominal interest rate rises by more than the increase in inflation, implying that the real yield rises by 43 basis points per percentage point of inflation.

In this case, inflation actually *raises* the *real* yield that pension funds can earn. Although this effect also reduces the inflation penalty on household investors, the impact of inflation on the *gap* between the household return (r_h) and the pension return (r_p) is increased. Since $r_p=i-\pi$ while $r_h=(1-\theta)i-\pi$, $d(r_p-r_h)/d\pi=\theta(di/d\pi)$. With $di/d\pi=1.43$ and $\theta=0.3$, the real yield spread rises by more than 40 basis points per percentage point of inflation.

In an all-debt economy, the truth would probably lie somewhere between the Fisherian constant-real-yield assumption and the alternative positive effect of inflation on the real interest rate. Whatever the exact value, a higher rate of inflation probably raises the real return that pension funds can earn and, by widening the gap between household and pension yields, encourages a greater role for pensions in private saving.

C. *An Economy with Debt and Equity*

In a more general economy with both debt and equity capital, the effect of inflation on pension funds is essentially a mixture of the results of the two simpler

[4] Although Irving Fisher's theoretical support for the independence of inflation and the real interest rate is no longer valid in an economy with a complex tax structure, the combination of tax rules and government debt policy can achieve an approximately constant real rate of return; see my 1980c article.

[5] This ignores the influence of noncorporate borrowers, government debt policy, and international capital flows.

[6] This is a special case of the result derived in my article with Jerry Green and Eytan Sheshinski.

[7] This is a linear approximation and also subsumes the excess taxation of nominal investory profits.

analyses.[8] A higher rate of inflation reduces the real net rate of return on equity for both households and pensions, but the reduction for pensions is less.[9] The real return on debt remains unchanged for pensions, but falls substantially for households ($dr_h/d\pi = -\theta$). Thus inflation slightly lowers the overall yield on pension funds, but, by widening the yield differential in favor of pensions, induces a switch in private saving in favor of pensions.

The relative yields on debt and equity move in opposite directions for households and pensions. For pensions, the real yield on debt is maintained while the real yield on equities fall slightly. For households, the real yield on debt falls sharply while the real yield on equities falls by less. Households thus sell debt to pension funds and hold more equity directly.

On balance, it is clear that a positive expected rate of inflation would not be a problem for the private pension system. The real net return earned by pension funds is relatively insensitive to inflation and might actually increase. In any case, the incentive to save through pensions would be strengthened by the increased differential between the yields on direct saving and pension saving.

II. Inflation and Pension Benefits

The typical private pension plan provides a retiree with a fixed nominal benefit based on his average earnings in the immediate preretirement years and on his number of years as an employee of the firm. Retirees of course find that the real value of their monthly pension benefit is continually reduced by the rising price level.

The switch from nominal pensions to indexed pensions would require either a reduction in the starting value of the pension or a reallocation of employees' lifetime incomes from wages to pensions. The amounts are significant; for any given benefit, indexing a single life annuity for a man age 65 with a 6 percent inflation rate would raise the cost of the pension by approximately 50 percent. This could be financed by a one-third reduction in the starting value of the pension, or roughly a 5 to 10 percent decrease in wages, or some combination of the two.

Although employees and firms have not yet faced and resolved this choice, an increasing use of (partially) indexed pensions is likely to evolve in future years. Some reduction in the starting value of the pension would probably be used, especially for long-service employees, to avoid overpensioning.[10] But on balance the move to indexing is likely to increase the total size of pension saving.

III. Uncertain Inflation

The uncertainty of inflation influences the optimal extent of pension indexing and the likely composition of pension assets. It is important to distinguish unanticipated changes in the *prive level* from unanticipated changes in the *expected future inflation rate*. A one-time rise in the price level lowers the real value of nominal assets but leaves the real value of equities unchanged. (see my 1980a article.) In contrast, a rise in the expected future rate of inflation leaves the value of short-term bills unchanged, but

[8] This assumes that pension funds behave as separate entities. Fischer Black and Irwin Tepper have noted that, if pension fund decisions are made as part of the overall corporate financial strategy, it is optimal for the pension fund to hold only debt while the rest of the corporate financial structure is adjusted to achieve the desired debt-equity mix for the corporation-cum-pension as a whole.

[9] My 1980b article presents explicit calculations with realistic values that imply that the equilibrium real net yields on equity (with a constant $f'=.112$) fall from .0966 for pensions and .0770 for households with no inflation to .0861 for pensions and .0648 for households with a 6 percent inflation.

[10] An employee with 30 years service in a firm with a pension plan now frequently receives a pension equal to 60 percent of his final five-year average earnings. For a retiree with average earnings, Social Security adds another 46 percent of final year's pay plus 23 percent more for a dependent spouse. The total easily reaches or exceeds 100 percent of real lifetime earnings at a time when family responsibilities, mortgage payments, and the like are relatively low. With an unindexed pension, this overpensioning is quickly offset by inflation.

lowers the nominal (and therefore the real) values of bonds and stocks.[11]

Without indexing, the vested pension obligations are nominal long-term liabilities of the firm. The firm can hedge these liabilities by holding long-term bonds. Of course, firms may nevertheless invest in equities because they believe that the equity yield is high enough to compensate for the reduced hedging.[12] But, since the extra risk of equity investment is borne by the firm's shareholders, the employees who participate in the pension plan should earn an implicit nominal return on their forgone wages that is only equal to the nominal return on riskless bonds.

A fully indexed pension would make all pension obligations real. Long-term bonds are clearly an inappropriate investment for funding such real obligations. Stocks can provide a hedge against price level uncertainty, but only by accepting substantial general uncertainty. Zvi Bodie has emphasized that a portfolio with a minimum-variance real return would be invested almost completely in short-term debt (with a small amount in commodity futures) and that the expected return on such a portfolio is approximately zero. If employees are so risk averse that they choose a fully indexed pension, the implicit real return that they earn on forgone wages should therefore also be approximately zero. Again, firms may invest in equities, but the shareholders rather than the pensioners should receive any extra yield in return for bearing that risk.

If employees choose only a partially indexed pension, that is, one in which benefits rise less than one-for-one with the price level or in which benefits depend on the return on the pension fund assets, the firm can invest in a way that permits giving a higher return to pension participants while compensating shareholders for any additional risk that they bear. The optimal extent of pension indexing depends on the risk aversion of employees and the cost, in terms of the reduction in the expected yield, of investing pension assets to produce a constant real return.

As Paul Samuelson noted years ago, an unfunded Social Security program can provide an annuity with an implicit real rate of return equal to the real growth rate of the economy, probably about 3 percent a year over the next decade or longer. Although 3 percent is substantially less than the real return of more than 10 percent that the nation as a whole earns on additions to the stock of plant and equipment (see my paper with James Poterba), the pressure to substitute unfunded Social Security benefits for private pensions (or vice versa) is likely to depend on the real *after-tax* yield that partly indexed pensions can offer and on the associated risk. If employees were completely risk averse, the low 3 percent yield on Social Security would look good in comparison to Bodie's zero yield on a minimum-variance real return portfolio. But if employees are willing to accept the risk inherent in a partially indexed pension, they can expect to receive an implicit yield that is much greater than 3 percent.[13]

In summary, the form and funding of private pensions will probably change in the coming decade if inflation continues at recent levels but, unless employees become much more risk averse, private pensions are likely to continue to finance a growing share of retirement consumption.

[11] See my 1980a article for an explanation of why stocks respond in this way to changes in the price level and the rate of inflation.

[12] For the reasons given by Black and Tepper, this equity investment might be done in the corporation itself with offsetting corporate borrowing. I shall not, in the remaining paragraphs, distinguish investment in the firm from investment in the pension portfolio.

[13] This yield will of course depend on both the performance of the economy and the changes in tax laws. In particular, indexed depreciation would reduce the riskness of equity investment (by reducing the sensitivity of share prices to the expected rate of inflation) and increase its expected return.

REFERENCES

F. Black, "The Tax Advantages of Pension Fund Investments in Bonds," Nat. Bur. Econ. Res. work. paper no. 533, 1980.

Z. Bodie, "Purchasing Power Annuities:

Financial Innovation for Stable Real Retirement Income in an Inflationary Environment," Nat. Bur. Econ. Res. work. paper no. 442, 1980.

J. Bulow, "Analysis of Pension Funding Under ERISA," Nat. Bur. Econ. Res. work. paper no. 402, 1979.

M. Feldstein, (1980a) "Inflation and the Stock Market," *Amer. Econ. Rev.*, Dec. 1980, *70*, 839–47.

______, (1980b) "Inflation, Tax Rules and the Stock Market," *J. Monet. Econ.*, July 1980, *6*, 309–31.

______, (1980c) "Fiscal Policies, Inflation, and Capital Formation," *Amer. Econ. Rev.*, Sept. 1980, *70*, 636–50.

______, **J. Green, and E. Sheshinski,** "Inflation and Taxes in a Growing Economy with Debt and Equity Finance," *J. Polit. Econ.*, Apr. 1978, *86*, S53–S70.

______ **and J. Poterba,** "State and Local Taxes and the Rate of Return on Non-Financial Corporate Capital," Nat. Bur. Econ. Res. work. paper no. 508R, 1980.

Irving Fisher, *The Theory of Interest*, New York 1930.

P. A. Samuelson, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *J. Polit. Econ.*, Dec. 1958, *66*, 467–82.

I. Tepper "Taxation and Corporate Pension Policy," mimeo., 1980.

Inflation, the Stock Market, and Owner-Occupied Housing

By LAWRENCE H. SUMMERS*

The past decade has witnessed major revaluations of the principal forms of capital held in the American economy. The real market price of corporate capital as reflected in the stock market declined by 45 percent between 1965 and 1980. During the same period, the real price of owner-occupied housing increased by 34 percent.[1] These capital gains and losses have had a substantial impact on the composition of wealth. In 1965, the market value of corporate capital exceeded that of owner-occupied housing by nearly 30 percent, yet by the end of 1979, the value of owner-occupied housing was almost twice the value of corporate capital. Since the market valuation of existing capital assets is a key signal guiding investment decisions, these revaluations also have important implications for economic performance.

This paper suggests that to a large extent, the increases in the value of housing and decreases in the value of corporate capital may have a common explanation, the interaction of inflation and a nonindexed tax system. The acceleration of inflation has sharply increased the effective rate of taxation of corporate capital income, while reducing the effective taxation of owner-occupied housing. These changes have been capitalized in the form of changing asset prices. In the long run, they will lead to significant changes in the size and composition of the capital stock.

The first section of the paper describes in more detail the nonneutralities caused by inflation. A simple model showing how inflation and taxation interact to determine asset prices is presented in the second section. The third section presents some crude empirical tests suggesting that increases in the expected rate of inflation may account for a significant part of the asset price changes which have been observed. A final section concludes the paper by commenting on some implications of the results.

I. The Effect of Inflation on Taxation

Inflation dramatically alters the relative tax treatment of corporate and noncorporate capital. The effective tax rate on corporate capital income is increased because of nonneutralities at both the corporate and individual level. At the corporate level, taxable profits are overstated because of historic cost depreciation and *FIFO* inventory accounting. At the individual level, the taxation of nominal as well as real capital gains significantly increases effective tax rates. An additional minor effect is the increase in marginal tax rates which occurs as inflation pushes taxpayers into higher brackets.

None of these features impact on the taxation of owner-occupied housing. The imputed rent is untaxed, and capital gains on real estate largely escape taxation because of roll over provisions, deferral, and the exemption for those over 55. Indeed, it is frequently argued that inflation actually reduces the tax burden on owner-occupied housing because of the deductibility of

*Assistant professor of economics, Massachusetts Institute of Technology, and National Bureau of Economic Research. I am indebted to Andrei Shleifer for valuable research assistance and to Patric Hendershott for helpful comments. This paper is based on research described in my 1980a, b papers where a much more detailed treatment of these issues is presented. The interaction of inflation and taxes in determining stock market and housing prices has previously been discussed by Martin Feldstein (1980) and Hendershott.

[1] The real price of corporate capital is measured as James Tobin's *q* ratio of the stock market value of equity plus the market value of debt divided by the capital stock. The data come from the National Balance Sheets prepared by the Federal Reserve Board. The real price of housing is calculated as the *GNP* deflator for single family construction divided by the personal consumption deflator.

TABLE 1

Year	Effective Tax Rate on Corporate Capital (1)	Real Net Return on Corporate Capital (2)	Effective Tax Rate on Owner-Occupied Housing (3)	Real Net Return on Owner-Occupied Housing (4)
1960	66.5	3.5	.32	5.0
1965	55.1	6.5	.37	4.5
1970	70.5	2.8	.40	4.3
1975	72.4	2.4	.34	4.4
1979	74.5	2.3	.32	4.3

Source: Cols. (1) and (2) are taken from Feldstein and Poterba. Cols. (3) and (4) are computed using a rental value series derived from *NIPA* estimates of imputed rent and a residential capital stock (owner-occupied) series provided by the Bureau of Economic Analysis. The effective property tax rate series, computed for use in the *MPS* model, was kindly made available by Franco Modigliani. In calculating the net rate of return depreciation is estimated at 1.5 percent per year. Maintenance costs are derived from the estimates in the FHA *Series Data Handbook*.

nominal rather than real mortgage interest payments. This argument is somewhat misleading, since nominal interest payments are also deductible on loans taken out to finance other types of investments. Thus, the deductibility of nominal interest does not encourage owner occupied housing vis-à-vis other forms of investment. Only if households find home ownership relaxes capital market constraints will the deductibility of nominal interest payments be relevant.[2]

The importance of taxation and its interactions with inflation in understanding the valuation and accumulation of corporate and housing capital is illustrated in Table 1. The table shows clearly the very heavy tax burden imposed on corporate capital. In 1979, corporate taxes at the federal and state levels combined with individual taxes on capital gains, dividends, and interest income captured over three-quarters of the capital income generated in the nonfinancial corporate sector. In contrast, property taxes impose only about a 30 percent burden on owner-occupied housing.

Table 1 also demonstrates that inflation has a large impact on the relative taxation of corporate capital and owner occupied housing. Despite the liberalization of depreciation allowances, increases in the investment tax credit and reduction in the corporate tax rate which have occurred since 1965, inflation has increased the effective tax rate on corporate capital from 55.1 percent to 74.5 percent. Since taxes capture the bulk of the return to corporate capital, increases in the tax rate have a very pronounced effect on after tax profits. If the tax rate in 1979 had been the same as it was in 1965, after-tax returns to the owners of corporate capital would have been almost twice as great.

Because the tax treatment of housing is inflation neutral, the effective tax rate has been almost constant between 1965 and the present. Thus, the interaction of inflation and taxation has very substantially altered the relative return on these assets. The next section develops a simple model of how these large changes affect the valuation and accumulation of corporate capital and housing.

II. The Determinants of Capital Market Equilibrium

It is useful to begin by assuming that inflation does not affect the real net of tax return required by investors. The determina-

[2] The fact that mortgage debt is the principal or only financial liability of many households suggests that housing investments are easier to borrow against than other alternatives. This may be because of their risk characteristics or of institutional constraints. These "leverage effects" are neglected below creating a presumption that the effect of the interaction of inflation and taxes is underestimated.

tion of the required real rate of return is considered below. In treating the corporate sector it is assumed that all new investment is financed through new equity issues, and that labor is supplied inelastically. I summarize the complex provisions of the tax code by assuming that the total tax rate on corporate capital depends linearly on the rate of inflation. These assumptions imply that dividends are given by

$$Div = F_k(1-\tau)K - \lambda\pi K \tag{1}$$

where F_k is the net marginal product of capital, τ is the tax rate on real corporate income, and λ reflects the nonindexation of the tax system.[3]

In order to induce investors to hold equity, it is necessary that the dividend yield plus the expected capital gain equal the required return on corporate assets; i.e.,

$$\frac{Div}{qK} + \frac{\dot{q}}{q} = \rho \tag{2}$$

where q is the market price of a unit of capital and ρ is the required rate of return. For simplicity, it is assumed below that investors have myopic expectations so that there are no expected capital gains or losses.[4] In this case, (2) implies that

$$q = \frac{F_k(1-\tau) - \lambda\pi}{\rho} \tag{3}$$

The evolution of the system depends on movements in the capital stock. These are assumed to be governed by a "Tobin's q" investment equation of the form:

$$\dot{K} = I(q) \quad I(1) = 0 \tag{4}$$

This equation indicates that the rate of net investment depends on the ratio of the market value of capital to its replacement cost. Equations (3) and (4) are sufficient to analyze the response of the corporate sector to a change in the rate of inflation.

Note that the initial impact of a change in the expected rate of inflation is to reduce the value of the stock market. The subsequent higher expected rate of inflation is correlated with an increasing value of q. Thus the theory predicts sharply different effects of expected and unexpected inflation. The decreases in the real price of corporate capital following increases in the rate of inflation are eliminated as the quantity of capital declines restoring the post-tax marginal product of capital to its initial level.

The housing sector can be modelled in an exactly parallel fashion as described in James Poterba, and in my 1981 paper. The market for the existing stock of housing must clear at each instant. Under the maintained assumption of myopic expectations, this implies that

$$R(H)/p_H = \rho + \theta \tag{5}$$

where $R(H)$ is the inverse demand curve for the rental services of housing, p_H is the price of housing in terms of the consumption good, ρ is the required real rate of return on capital, and θ is the net property tax rate plus any constant risk differential between the required rates of return on housing and corporate capital. The accumulation of housing is governed by a supply equation of the form:

$$\dot{H} = h(p_H) \qquad h(1) = 0 \tag{6}$$

These two equations can be used to analyze the dynamics of housing prices and investment. Note that unless inflation affects ρ, the required rate of return, it has no effect on the real price or accumulation of housing capital. If account were taken of the "leverage effects" focused on by Poterba, increases in inflation would raise the real price of housing, even if ρ remained constant.

So far, the analysis has been partial equilibrium in that ρ, the required real rate of return on capital has been determined exogenously. In order to examine the impact of

[3] My 1980a paper considers a model with a much more realistic tax system, and more realistic assumptions about financial policy. The qualitative conclusions of the current analysis continue to hold.

[4] The consequences of relaxing this assumption and assuming perfect foresight are discussed below and in my 1980b paper.

inflation on asset prices, it is necessary to model explicitly its impact on ρ. This is done by imposing the requirement that at every instant, aggregate demand and supply are equal; i.e.,

$$(7)\quad C(YL, qK+p_H H)+I(q)K + p_H h(p_H)+G=F(K,L)+R(H)H$$

where $C(YL, qK+p_H H)$ is a life cycle consumption function depending on labor income and wealth, and the right-hand side is aggregate supply which equals the sum of housing services and other produced output. Using (3) and (5), equation (7) can be written for given values of the exogenous variables as

$$(8)\quad AD(K,H,\rho)=AS(K,H)$$

The model can be solved in two different ways. The short-run effect of inflation on asset valuation is calculated by first solving (8) treating K and H as predetermined, and then using (3) and (5) to find the asset prices. It is easy to verify that $\partial\rho/\partial\pi<0$. This is because the required rate of return must fall to restore equilibrium after inflation reduces corporate investment. It is through this decline in ρ that an inflation shock is transmitted to the housing sector. Thus it is also clear that $\partial q/\partial\pi<0$ and $\partial p_H/\partial\pi>0$. These price changes spur changes in the composition of the capital stock. The steady-state solution of the model is found by setting $q=p_H=1$. It can be shown that inflation unambiguously reduces the corporate stock, and for plausible parameter values raises the housing stock.

The quantitative importance of these effects can be illustrated by inserting plausible parameter values into the model. The key parameter choice is the value of $\lambda\pi$ in equation (3). Following the evidence in Table 1, $\lambda\pi$ is taken to equal half the value of the after tax real return to capital in the absence of inflation. Space constraints preclude a discussion of the remaining parameter choices, but variations in them have little effect. The model implies that the short-run effect of an inflation-induced increase in taxes of the size witnessed in the United States would be to reduce the value of q by 35 percent and to raise the value of houses by 29 percent. The long-run effect, after housing and stock prices have returned to their equilibrium levels, is to reduce the corporate capital stock by one-third and to increase the housing stock by 2 percent. The housing stock increase is so small because of the reductions in labor income which accompany the declining corporate capital stock.

These results may overstate somewhat the true adjustment in prices to an inflation shock because of the myopic expectations assumption. Rational investors would recognize that capital gains and losses would take place as the corporate capital stock falls and the housing stock rises. These expectations will attenuate the initial price jumps. However, available evidence (see my 1980 paper) suggests that the response of investment to shocks is very sluggish so that the overstatement is not large. In my 1981 paper, it is shown that, even though the return of asset prices to their equilibrium values can be predicted, there will be no predictable excess returns for investors in either asset.

The predictions of this simple model accord with the realities of recent years. A variety of other explanations for the decline in stock prices and increases in the value of housing have been suggested. Most account for one phenomenon or the other. Indeed, consideration of both results together tends to discredit several popular explanations. For example, the Modigliani-Cohn interest illusion view would predict that housing prices should also have fallen. Presumably, middle-income mortgage borrowers are more likely to be duped by high nominal interest rates than the more financially sophisticated investors who dominate the stock market. The theory advanced here has the substantial virtue of providing a unified explanation of these phenomena. The next section presents some empirical evidence tending to support the hypothesis that increases in expected inflation account for a large fraction of the change in the relative values of corporate and housing capital.

III. Empirical Estimates

The model developed in the preceeding section has the clear implication that in the short run an increase in the permanent expected rate of inflation should increase the market price of housing and reduce the value of the stock market. This proposition is tested by regressing the excess return on the stock market and on houses against the change in the permanent expected rate of inflation. That is, the equations estimated are of the form:

$$R_t = \alpha + \beta \Delta \pi_t^e + u_t \tag{9}$$

The excess return on the stock market is defined as the sum of capital gains and dividends as a percent of beginning of period market value less the beginning of period Treasury bill rate. Since imputed rents are difficult to measure, they are ignored in calculating the excess return on owner occupied housing. The return is calculated as the appreciation in the price deflator for one-family structures, prepared by the BEA, less the beginning of period Treasury bill rates.

The argument made here holds that returns should depend on changes in the expected rate of inflation. The expected rate of inflation is estimated using the "rolling *ARMA*" approach employed in my article with Feldstein. To estimate the expected permanent rate of inflation in year t, an *ARMA* process is fitted to the preceding 10 years data on inflation. It is then used to forecast the next 10 years' inflation. The expected permanent rate is taken to be a discounted weighted average of these forecasts.[5]

The estimated stock market equation using quarterly data for the 1958–78 period is

$$\begin{array}{llll} R_t = & .126 & -7.61\Delta\pi^e & R^2 = .16 \\ & (.029) & (1.61) & DW = 1.97 \end{array}$$

Increases in the expected rate of inflation show a strong negative relationship to security returns, as theory would predict. The results imply that a 1 percent increase in the expected rate of inflation reduces the value of the stock market by 7.61 percent. This means that an increase in the expected rate of inflation from 3 percent in the late 1960's to 8 percent today can account for almost all of the stock market's real decline over the recent period. This result is quite robust. Including additional variables, such as the expected rate of inflation, unexpected profits, and measures of risk do not significantly alter the estimate of the expected impact of inflation. Using alternative proxies for the safe asset such as the commercial paper rate also has a negligible effect on the results. The equation is also insensitive to the choice of sample period.

The corresponding equation for the return on owner occupied housing is

$$\begin{array}{lll} R_t = .004 + & 1.68\Delta\pi^e & R^2 = .02 \\ & (1.08) & DW = 1.3 \end{array}$$

It also suggests the importance of unexpected inflation, though the effects are much smaller than in the stock market. This may be due to the numerous financial market imperfections which have made the housing market illiquid during periods of high inflation. During such periods transactions prices may not reflect real values since investors are constrained. When the equation was reestimated omitting credit crunch periods, the result was[6]

$$\begin{array}{lll} R_t = .003 + & 3.12\Delta\pi^e & R^2 = .09 \\ & (1.21) & DW = 1.7 \end{array}$$

This suggests that recent financial market innovations, including the introduction of variable rate mortgages and the relaxation of regulation Q, are likely to increase the sensitivity of housing returns to inflation.

[5] The *ARMA* process is assumed to be first-order autoregressive, first-order moving average. The rate of inflation is calculated from the consumption price deflator. In calculating the long-term rate of inflation, an 8 percent discount rate is used.

[6] Credit crunch periods were identified based on the crunch dummy in the *MPS* model. They include 60:1–60:3, 66:3–67:2, 69:2–70:1 and 73:3–75:1.

Future research will attempt to decompose movements in asset prices into fractions due to innovations in inflation, tax laws, and other variables. Until this is done, the results here must be viewed as indicative, but not demonstrative, of the importance of the inflation-tax interactions described. Most other hypotheses, which have been advanced for the changes in capital asset prices, do not have the implication that changes in the expected rate of inflation should have the systematic effects found here.

IV. Conclusions

This paper shows that the nonneutral effect of inflation on our tax system can account for much of the increase in the value of owner-occupied housing and reduction in the value of the stock market which has occurred in recent years. The theory suggests that, in the long run, high rates of inflation are likely to shift substantially the composition of the capital stock. Movements in this direction have already been observed. The share of net business fixed investment in *GNP* has fallen from 3.1 percent in the 1960's to 2.4 percent in the 1970's. The share of net housing investment has declined only slightly from 2.5 percent to 2.4 percent. These changes are likely to have significant implications for long-term economic growth.

REFERENCES

M. Feldstein, "Inflation Tax Rules and the Stock Market," *J. Monet. Econ.*, July 1980, *6*, 309–31.

______ **and J. Poterba,** "State and Local Taxes and the Rate of Return on Non-Financial Corporate Capital," Nat. Bur. Econ. Res. working paper no. 508, 1980.

______ **and L. Summers,** "Inflation, Tax Rules, and the Long Term Interest Rate," *Brookings Papers*, Washington 1978, *1*, 61–99.

______ **and** ______, "Inflation and the Taxation of Capital Income in the Corporate Sector," *Nat. Tax. J.*, Dec. 1979, *32*, 445–70.

P. Hendershott, "The Decline in Aggregate Share Values: Inflation and Taxation of the Returns from Equities and Owner Occupied Housing," Nat. Bur. Econ. Res. working paper no. 370, 1977.

J. Kearl, "Inflation Induced Distortions in the Real Economy, An Econometric and Simulation Study of Housing and Mortgage Innovations," unpublished doctoral dissertation, MIT 1975.

F. Modigliani and R. Cohn, "Inflation, Rational Valuation and the Market," *Financial Analysts J.*, Mar.-Apr. 1979.

J. Poterba, "Inflation, Income Taxes and Owner-Occupied Housing," Nat. Bur. Econ. Res. working paper, 1980.

L. Summers, "Inflation, Taxation and Corporate Investment" mimeo., 1980.

______, "Capital Taxation in a General Equilibrium Perfect Foresight Growth Model," mimeo., 1981.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
NINETY-THIRD
ANNUAL
MEETING

DENVER, COLORADO
SEPTEMBER 5–7, 1980

Minutes of the Annual Meeting
Denver, Colorado
September 6, 1980

The Ninety-Third Annual Meeting of the American Economic Association was called to order by President Moses Abramovitz at 9:42 P.M., September 6, 1980, in the Imperial Ballroom of the Fairmont Hotel. The minutes of the meeting of December 29, 1979 were approved as published in the *American Economic Review Proceedings*, May 1980, pages 427–30.

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), the Managing Editor of the *American Economic Review* (George Borts), the Managing Editor of the *Journal of Economic Literature* (Mark Perlman), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were distributed at the meeting. (See their reports published elsewhere in this issue.) Werner Hasenberg asked if the recent agreement with Hertz was an indication of a trend toward commercialism. The Secretary responded that the Association avoided entangling commercial alliances unless it was felt that significant benefits accrued to the members. In this case, there appeared to be an opportunity to provide a service to the members at no cost. The agreement was not exclusive and could be terminated on thirty days' written notice. The Secretary did not believe the Executive Committee intended to establish a trend toward commercialism when they approved the agreement. William Vickrey asked about extending the meetings so that fewer simultaneous sessions would be necessary. The Secretary responded that more days of meetings would make his life easier in planning the meetings but he doubted if the members would prefer such an extension. He pointed out that the affiliated societies plan sessions somewhat independently of the AEA and each other. The current policy is not to exclude new societies from ASSA nor to dictate the number of sessions they may organize. If the program continues to grow, the policy may have to be revised.

The Treasurer reported on two actions of the Executive Committee. The Committee voted to raise the base rate of membership dues to $30 effective January 1, 1981 if the amendment to the bylaws submitted to the membership is approved, or to $31.50 effective April 1, 1981 if it is not approved. Secondly, the Treasurer was authorized to raise subscription rates to nonmembers up to $120 after investigating the legal and tax ramifications.

After Borts' report, his last as Managing Editor of the *AER*, William Baumol moved, Elizabeth Bailey seconded, and it was VOTED unanimously to express on behalf of the Association a sense of deep gratitude and appreciation of George Borts' long and devoted efforts on behalf of our discipline in his role as editor of the *American Economic Review*.

After Perlman's report, his last as Managing Editor of the *JEL*, Robert Solow moved, William Vickrey seconded, and it was VOTED unanimously to express, on behalf of the Association, our great appreciation to Mark Perlman for his development of the *Journal of Economic Literature* into one of our disciplines' most useful journals during his tenure as its first editor.

Baumol then rose to express his thanks to members of the AEA's staff, especially Mary Winer, Barbara Weaver, Violet Sikes, Melissa Williams, and Norma Ayres, and to his associate Sue Anne Blackman, for their aid in helping him organize the 1980 program..."They are efficient, intelligent, and charming." The meeting responded with loud applause.

The Secretary presented the following resolutions, which were adopted unanimously:

> BE IT RESOLVED that this meeting record a special note of thanks to the members of the 1980 Allied Social Science Associations' Convention Committee, chaired by Thomas E. Davis,

for their hard work and efficient management of these meetings.

BE IT RESOLVED that this meeting commend William Baumol for planning a program of great interest and distinction.

The Chair then introduced William Baumol, the President-elect. There being no further business, the meeting adjourned at 10:22 P.M.

C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in Washington, D.C., March 21, 1980.

The first meeting of the 1980 Executive Committee was called to order at 9:15 A.M. on March 21, 1980 in the Conservatory Room of the Washington Hilton Hotel. The following members were present: Moses Abramovitz, presiding, Henry J. Aaron, William Baumol, George Borts, Carl Christ, Martin Feldstein, Rendigs Fels, C. Elton Hinshaw, H. Gregg Lewis, Robert Lucas, Mark Perlman, and Robert Solow. Leo Raskind was present as the Association's Counsel. Present as members of the Nominating Committee were Lawrence Klein (chair), Herman Daly, David Kendrick, Anna J. Schwartz, Lester Thurow, and Murray L. Weidenbaum. Present as guests for parts of the meeting were Marcus Alexis, Donald Brown, Franco Modigliani, Naomi Perlman, and Wilma St. John.

Minutes. The minutes of the meeting of December 27, 1979 were approved.

Report of the Secretary (Hinshaw). The Secretary reported that the schedule for future annual meetings is Denver, Colorado, September 5–7, 1980; Washington, D.C., December 28–30, 1981; New York, December 28–30, 1982; San Francisco, December 28–30, 1983; Dallas, Texas, December 28–30, 1984; and New York, December 28–30, 1985. Because the 1980 meetings will occur early in the academic year, a separate placement meeting will be held in Dallas, Texas, December 28–30, 1980. Placement services will not be provided at the Denver meetings.

He reviewed the Association's policy of granting complimentary memberships (without the privilege of receiving publications) to all members sixty-five years of age or older who have retired from their professional activities, provided they have been members of the Association for twenty-five years or more. It was VOTED to modify the policy by increasing the age limit to sixty-eight years or older and decreasing the required time of membership to twenty years. Such complimentary memberships are not entitled to receive publications but have all other rights and privileges of regular members.

It was VOTED to conclude an agreement with Hertz whereby AEA members would be granted a 20 percent discount on car rental rates, subject to further investigation of other offers.

During the discussion of A. W. Coats' proposal to write a history of the Association, it was decided to appoint an *ad hoc* committee charged with establishing a general policy for handling the Association's archival material and determining rules of access to the archives. It was VOTED that the Executive Committee welcomes Coats' proposal and will cooperate by making the files available subject to the recommendations of the *ad hoc* committee concerning access. It was understood that the history will not be an official one, the Association will not contribute funds to support the project, but officers will write letters of support to prospective funding agencies.

Report of the Treasurer (Fels). The Treasurer reported that, thanks to an increase in the market value of the Association's portfolio toward the end of the year, the surplus for 1979 was $141 thousand, instead of the $108 thousand anticipated in December. The projected deficit for 1980 is now $35 thousand. The dues increase is still needed in 1981 despite the more favorable outlook for 1980. Upon the Treasurer's recommendation to invest some funds in a money market mutual fund, the following resolution was VOTED:

> RESOLVED that this corporation enter into an agreement with Stein Roe Cash Reserves, Inc. authorizing Stein Roe Cash Reserves, Inc. to honor any telegraphic, telephonic, or written requests believed by it to be authentic for redemption of shares in the corporation's account in accordance with the purchase application form herewith submitted.

The Treasurer was asked to send a report on the performance of Stein Roe and Farnham's management of the portfolio to interested members of the Executive Committee and that such performance analyses should be a part of future Finance Committee reports.

Report of the Editor of the American Economic Review (Borts). Upon the recommendation of the editor, it was VOTED to appoint the following persons to the Board of Editors: Albert Ando, Herschel I. Grossman, Peter W. Howitt, Anne O. Krueger, James P. Smith, and Robert D. Willig.

Report of the Editor of the Journal of Economic Literature (Perlman). The editor reviewed the progress in and plans for the transfer of the journal to the new editor. He announced that the 1976 and possibly the 1977 *Index of Economic Articles* would be published this year.

Centennial Celebration (Perlman). The special exploratory Committee on the Association's Centennial Celebration proposed a special one-day session of the 1985 annual meeting be held in Washington, D.C. There should be several papers or panel discussions retro- and prospectively viewing the various subdisciplines—a century before (1885) and a century afterwards (2085). These papers should be incorporated in the *Proceedings* or in a handbook issue. There should be an informal banquet honoring past officers, Foreign Fellows, Distinguished Fellows, and representatives of foreign societies. It was decided to have an extra half-day at the 1985 meetings in New York to devote special attention to the Centennial celebration.

International Economic Association (Modigliani). It was VOTED to indicate strong interest in further exploration of a joint meeting of the American Economic and International Economic Associations in December 1985; pending additional investigation of the feasibility of such a meeting, the Association will issue an invitation to the IEA to meet with us in New York.

Encyclopaedia (Perlman). Perlman indicated that he thought a 5-million word, 8-volume encyclopaedia would have a reasonably strong market. In addition to the encyclopaedia, he proposed a set of about eight "Handbooks" of economics subfields, varying in size from 200,000 to 400,000 words; they would contain selected material (updated) drawn from the encyclopaedia. A third product could possibly be an etymological dictionary of economics, which is a necessary "inhouse" input in the production of the encyclopaedia and the handbooks.

The following motion was VOTED: The Executive Committee approves in principle the preparation of an "Encyclopaedia of Economics," provided that a satisfactory detailed scheme of administration and production can be worked out. It asks the President to appoint a committee to study alternatives and to report to the September 1980 meeting of the Executive Committee with an evaluation of alternative schemes of administration and financing, a recommendation not to proceed or to proceed with a particular scheme, and if the latter, with tentative nominations for the members of an Encyclopaedia Council.

1980 Program (Baumol). Baumol reported that the 1980 program is in place. The only innovation is that he is planning four sessions per day instead of the usual three.

Nominating Committee (Klein). Klein reported the following nominees for offices in the 1980 election: Vice President (two to be chosen), Otto Eckstein, Leonid Hurwicz, Dale Jorgenson, and Alice Rivlin; Executive Committee members (two to be chosen), Elizabeth Bailey, Robert J. Gordon, Jacob Mincer, and Sherwin Rosen. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, chose as nominee for President-elect, Gardner Ackley, and as Distinguished Fellows, Solomon Fabricant and Charles Kindleberger.

Committee on the Status of Minority Groups in the Economics Profession (Alexis). Alexis reported that the 1980 summer program will be moved to Yale with the understanding that, if the Sloan Foundation continues to finance the project, Yale will keep the program for 1981 and 1982. Yale accepted the program subject to Donald Brown being appointed the Director, courses being taught exclusively by Yale faculty, and an additional course in econometrics being offered.

The Executive Committee approved the change. Donald Brown reported a slight upgrading of the prerequisites for entry. He now recommends one year of calculus.

It was VOTED to appropriate $10,000 for the 1980 program with the understanding that the money will not be spent if other funds become available.

Committee on the Status of Women in the Economics Profession (Bailey). Bailey reported that there are 139 women faculty in the sixty-five departments of the Chairperson's Group. However, forty of them have no tenured women; eighteen have only one tenured woman and seven have two tenured women. Moreover, seventeen of the sixty-five departments have no women, tenured or nontenured.

It was VOTED to appropriate $10,500 for the work of the Committee.

Terms of the Secretary and Treasurer (Abramovitz). It was VOTED to extend the terms of both the Secretary and Treasurer to December 31, 1984.

Other Business. It was decided to hold the 1981 spring meeting of the Executive Committee in Chicago.

The meeting was adjourned at 5:26 P.M.

Minutes of the Meeting of the Executive Committee in Denver, Colorado, September 4, 1980.

The second meeting of the 1980 Executive Committee was called to order at 10:06 A.M. on September 4, 1980 in the Denver Hilton Hotel, Denver, Colorado. The following members were present: Moses Abramovitz (presiding), Henry J. Aaron, William Baumol, George Borts, Carl Christ, Robert Clower, Martin Feldstein, Rendigs Fels, C. Elton Hinshaw, Robert Lucas, Mark Perlman, Robert Solow, Marina Whitman. Also present were Leo Raskind, counsel and Gardner Ackley, nominee for President-elect. Present as guests for parts of the meeting were Marcus Alexis and Elizabeth Bailey.

Minutes. The minutes of the March 21, 1980 meeting were approved as written.

Report of the Secretary (Hinshaw). The Secretary reported that the 1981 annual meetings are scheduled for Washington, D.C. on December 28–30 with headquarters at the Washington Hilton Hotel. Subsequent meetings are scheduled for New York, December 28–30, 1982; San Francisco, December 28–30, 1983; Dallas, December 28–30, 1984; and New York, December 28–30, 1985. Because the Denver meetings occur early in the academic year, a separate placement meeting is planned for December 28–30, 1980 in Dallas.

Z. A. Silberston, Secretary of the Royal Economic Society, has informed the Secretary that the Society has decided to pursue the plan for making bulk purchases of the *Journal of Economic Literature* for distribution to their members. We are in the process of working out the arrangements.

The Copyright Clearance Center, a nonprofit organization, operates a nationwide service that collects royalty fees and conveys permission to photocopy to users whenever their photocopying needs exceed the exemptions provided under the copyright law. It was decided not to register the AEA journals with the Center at this time.

After verifying that others did not offer better discounts than those proposed by Hertz, the Secretary signed an agreement whereby AEA members are entitled to a discount on car rentals from Hertz. The agreement was effective as of August 1, 1980, and will remain in effect until cancelled by either party upon thirty days' notice. Members have been sent AEA/Hertz identification cards.

The Secretary also reported that he plans to produce a directory of members in 1981. He expects to enter into a contract with a printer this fall, send out the questionnaires in the spring, and publish the directory as the second December 1981 issue of the *American Economic Review*.

Report of the Treasurer (Fels). Although the Treasurer circulated a preliminary 1981 budget that was virtually in balance, he requested that formal consideration of the budget be postponed until the March meeting. Sometime after Thanksgiving when 1980 revenues and expenses can be more accurately projected, he will circulate a proposed 1981 budget. He reported that the Budget Committee recommended to the Executive Committee that he be authorized to increase the subscription rate to nonmembers up to $100; that the submission fee for the *AER* be increased to $25; and that the base rate for membership dues be increased to $30 effective January 1, 1981 if the amendment to the bylaws is approved, or to $31 effective April 1, 1981 if it is not. It was moved to accept the recommendation to authorize the Treasurer to raise the subscription rate up to $100 after investigation of the tax and legal ramifications. The motion was amended to authorize the Treasurer to raise the rate up to $120 if in his judgment the subscription package should be expanded to include the annual *Index of Economic Articles*. The amendment was accepted, and it was VOTED to approve the amended motion. It was VOTED to raise the submission fee to the *AER* from $15 to $25. It was VOTED that (1) if the amendment to the bylaws is approved, the base rate for dues be raised to $30 effective January 1, 1981, and (2) if the amendment is disapproved it be raised to $31.50 effective April 1, 1981.

The Treasurer distributed information provided by Stein Roe & Farnham, investment counsel, on the performance of the Association's portfolio. The total return on the Association's portfolio during the past three years has exceeded the Dow Jones Industrial Average, Standard and Poors *500 Index* and the Lipper Composite for Balanced Funds. It was suggested that additional analysis include risk adjusted returns and comparisons with indexes other than those on the "Big Board."

Committee on the Status of Women in the Economics Profession (Bailey). Bailey reported that CSWEP is reorganizing itself, emphasizing its growing function as an umbrella organization for women involved in the regional economics associations. She distributed a draft report of an analysis of the CSWEP roster (the revised report appears elsewhere in this *Papers and Proceedings*) that discussed employment status, highest degree, publications, employment record, and fields of specialization. It was VOTED to make the 1978 directory tapes available to CSWEP.

Committee on the Status of Minorities in the Economics Profession (Alexis). Alexis, on behalf of the Committee, reported that the major activity continues to be the summer program. The program has been moved to Yale University where Don Brown now administers it. The 1980 program had twenty-eight students. The Sloan Foundation evaluation is nearing completion. Preliminary indications are that Sloan will continue to fund a major part of the program; the Federal Reserve System is increasing its level of support. It now appears that $10,000 from the AEA will be adequate. The Committee is now working with the National Economic Association to analyze what has been happening to minorities in the profession: number of economists, promotion through ranks, etc.

Report of the Editor of the American Economic Review (Borts). Borts made his last report as editor of the *AER*. (See his full report elsewhere in this *Papers and Proceedings*.) The new managing editor will be Robert W. Clower of the University of California-Los Angeles. The editorial office will move to UCLA January 1, 1981. Starting October 1, 1980, all new manuscripts will be processed at the new editorial office. The December 1980 issue will be produced from the Providence office. Subsequent issues will be produced by the Los Angeles office.

Report of the Editor of the Journal of Economic Literature (Perlman). Perlman made his last report as editor of the *JEL*. (See his full report elsewhere in this *Papers and Proceedings*.) The new managing editor will be Moses Abramovitz of Stanford University. Beginning with the June 1981 issue, the *JEL* will be edited in the Stanford editorial office. Perlman also reported that the 1976 *Index of Economic Articles* should be published in November of 1980, and the 1977 and 1978 *Indexes* should appear during 1981. Upon the recommendation of the new editor, it was VOTED to approve the appointments of David Laidler, Finis Welch, Roger Noll, Alan Blinder, John Monthias, and Michael Rothschild to the Board of Editors. It was proposed that the new editor run a periodic or annual survey of economic journals to determine the lags in the reviewing, rejecting, accepting, publishing process and to publish the results in the *JEL*. The proposal was referred to the Committee on Publications.

Report of the Director of Job Openings for Economists (Hinshaw). (See elsewhere in this *Papers and Proceedings* for his full report.)

Ad Hoc Committee on Encyclopaedia of Economics (Baumol). Baumol reported that the two most interested publishers were unwilling to fund the project. It is now apparent that the AEA will have to raise funds to finance the publication. It was moved that the Executive Committee approve with enthusiasm the proposed encyclopaedia project, that the Association apply to the National Endowment for the Humanities for a grant to help finance it, that (1) if the grant is not awarded, the project be dropped, and (2) if the grant is awarded, a subcommittee be appointed to raise the additional necessary funds. The motion was amended to read that if the grant was awarded, the Association would be willing to commit up to $200,000 (in 1981 dollars) per year for three years. A straw vote testing the Committee's degree of enthusiasm to support the project indicated a lack thereof. The motion was withdrawn without dissent.

1981 Program (Ackley). The 1981 Program Chair reported that he had started working on the program and had already received a large number of suggestions. He pointed out that a call for papers had been published in the September *AER*. The matter of publishing discussants' comments was discussed but no policy change was made. As in the past, the decision was left to the program chair.

International Economics Association (Abramovitz). The President reported on comments he had received from various persons who attended the Sixth World Congress of the IEA. It has been alleged that the local sponsoring agency appeared not to have been a proper economic association but more a political group; highly tendentious reports were fed to and published in the Mexican press; the program itself was of a low scientific level with a high political content; representatives of the Mexican government had intervened in the internal affairs of the IEA; and reaction of many of the participants was sufficiently vigorous that Victor Urquidi, the newly elected President of the Association, has asked to meet with the Executive Committee. It was VOTED that the President of the AEA be authorized to convey to the President of the IEA the Committee's grave concern about the conduct of the Sixth World Congress and be delegated the authority to withdraw the AEA from future participation in the IEA if such action became necessary to protect the integrity and reputation of the AEA.

There being no further business, the meeting was adjourned at 5:15 P.M.

C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1980

Annual Meetings. In 1981 the annual meetings will be held at the Washington Hilton Hotel in Washington, D.C. on December 28–30. The schedule for subsequent meetings is December 28–30, 1982 in New York, December 28–30, 1983 in San Francisco, December 28–30, 1984 in Dallas, and December 28–30, 1985 in New York.

Employment Services. For those meetings scheduled for December 28–30, employment services will be provided at the annual meeting beginning December 27.

The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision of the employment service provided at the annual meetings.

Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists,* and their professional obligation to list their openings.

Hertz Agreement. The Association has entered into an agreement with Hertz whereby our members are entitled to a 20 percent discount on car rentals within the United States (except Florida, Alaska, Hawaii, and Puerto Rico) and Japan on basic rates; a 10 percent discount in Florida, Alaska, Hawaii, Puerto Rico, Europe, Africa, the Middle East, Asia, the Pacific, and Latin America; and a 30 percent discount in Canada. The agreement was effective as of August 1, 1980.

1981 Directory. It is my current plan to publish a 1981 *Directory of Members* as a second December 1981 issue of the *American Economic Review.*

Membership. The total number of members and subscribers, shown in Table 1, reached an all-time high of 26,787 at the end of 1975. After declining for two years, the total increased in 1978, 1979, and again in 1980.

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1978	1979	1980
Class of Membership			
Annual	15,698	16,203	16,219
Junior	1,857	1,884	1,811
Life	389	388	383
Honorary	35	35	33
Family	307	315	331
Complementary	615	634	624
Total Members	18,901	19,459	19,401
Subscribers	6,893	6,963	7,094
Total Members and Subscribers	25,794	26,422	26,495

Permission to Reprint and Translate. Official permissions to quote from, reprint, or translate and reprint articles for the *American Economic Review* and the *Journal of Economic Literature* totaled 309 in 1980 compared to 262 in 1979. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

AEA Staff. I wish to take this opportunity to express my gratitude to the staff of the Secretary–Treasurer's office in Nashville. They are efficient, hard working, and dedicated, and handle the day-to-day operations of the Association. They do not attend the annual meetings and their names do not appear in our publications, but the Association depends heavily upon their talents. On your behalf and mine, I wish to thank them. They are Mary Winer, Administrative Director, Norma Ayres, Stephanie Baker, Ersye

Burns, Marcia McGee, Violet Sikes, Dale Wagner, Jacquelyn Woods, and Ettamene Byrd.

Committees and Representatives. Listed below are those who served the Association during 1980 as members of committees or representatives. The year in parenthesis indicates the final year of the term to which they have been appointed most recently. On behalf of the Association, I wish to thank them all for their services.

AD HOC COMMITTEE TO CONSIDER AND RECOMMEND CENTENNIAL ACTIVITIES
Irma Adelman
George Borts
Mark Perlman

AD HOC COMMITTEE TO DETERMINE POLICY FOR THE ASSOCIATION'S ARCHIVAL MATERIAL
William N. Parker, *Chair*
Michael Edelstein
Stanley Lebergott

AD HOC COMMITTEE ON ENCYLOPAEDIA OF ECONOMICS
William Baumol, *Chair*
Henry J. Aaron
Carl F. Christ

AD HOC ADVISORY COMMITTEE TO THE NATIONAL COMMISSION ON EMPLOYMENT AND UNEMPLOYMENT STATISTICS
Harold Watts, *Chair*
Orley Ashenfelter
Carolyn Shaw Bell
Charles C. Holt

AD HOC COMMITTEE ON PUBLISHING CONTRACTS
Martin Shubik, *Chair*
Peggy Heim
Leo Raskind
C. Elton Hinshaw, *ex officio*

AD HOC COMMITTEE TO STUDY SGE REPORT
Henry J. Aaron, *Chair*
George Jaszi
Hyman Kaitz

BUDGET COMMITTEE
Marina v. N. Whitman, *Chair* (1980)
Henry J. Aaron (1981)
Martin Feldstein (1982)
Rendigs Fels, *ex officio*
Moses Abramovitz, *ex officio*
William Baumol, *ex officio*

CENSUS ADVISORY COMMITTEE
Robert F. Lanzillotti, *Chair* (1981)
Otto Eckstein (1980)
Victor R. Fuchs (1980)
George L. Perry (1980)
Norman J. Simler (1980)
Lester C. Thurow (1980)
Barbara Bergmann (1981)
Martin H. David (1981)
Richard D. Karfunkle (1981)
William Niskanen (1981)
Carolyn Shaw Bell (1982)
Ronald L. Oaxaca (1982)
Thomas Sowell (1982)
Ann D. Witte (1982)
Arnold Zellner (1982)

COMMITTEE ON ECONOMIC EDUCATION
Allen C. Kelley, *Chair* (1982)
George Leland Bach (1980)
William E. Becker (1980)
Keith Lumsden (1980)
Karl E. Case (1981)
W. Lee Hansen (1981)
John Siegfried (1981)
Campbell R. McConnell (1982)
Rendigs Fels, *ex officio*

ECONOMICS INSTITUTE POLICY AND ADVISORY BOARD
Edwin S. Mills, *Chair* (1981)
Carlos F. Díaz-Alejandro (1980)
Raymond Vernon (1980)
Axel Leijonhufvud (1981)
Douglass C. North (1982)
Dwight Perkins (1982)
G. Edward Schuh (1982)
Bent Hansen (1983)
Louis Wells (1983)

COMMITTEE ON ELECTIONS
Ben Bolch, *Chair* (1980)
Gayle D. Riggs (1981)
C. Elton Hinshaw, *ex officio*

FINANCE COMMITTEE
Robert Eisner, *Chair* (1980)
James Lorie (1981)
Robert G. Dederick (1982)
Rendigs Fels, *ex officio*

COMMITTEE ON HONORARY MEMBERS
Leonid Hurwicz, *Chair* (1980)
Paul A. Samuelson (1980)
Hollis B. Chenery (1982)
Tibor Scitovsky (1982)
Hendrik S. Houthakker (1984)
George J. Stigler (1984)

COMMITTEE ON HONORS AND AWARDS
John Chipman, *Chair* (1980)
James W. McKie (1980)
Carl F. Christ (1982)
Anne O. Krueger (1982)
Dale T. Mortensen (1984)
Daniel McFadden (1984)
Oliver E. Williamson (1984)

NOMINATING COMMITTEE (1980)
Lawrence Klein, *Chair*
Herman E. Daly
David A. Kendrick
Sherwin Rosen
Anna J. Schwartz
Lester Thurow
Murray L. Weidenbaum

COMMITTEE ON POLITICAL DISCRIMINATION
Carl M. Stevens, *Chair* (1980)
Kenneth J. Arrow (1980)
John G. Gurley (1980)
Harold Barnett (1981)
Anne P. Carter (1981)
Martin Bronfenbrenner (1982)
Lester Thurow (1982)

COMMITTEE ON PUBLICATIONS
Robert Ferber, *Chair* (1981)
Martin Bronfenbrenner (1981)
Kenneth W. Leeson (1981)
Barbara Reagan (1981)
Edwin Burmeister (1982)
Peter A. Diamond (1982)
C. Elton Hinshaw, *ex officio*

COMMITTEE ON THE STATUS OF MINORITY GROUPS IN THE ECONOMICS PROFESSION
Marcus Alexis, *Chair* (1980)
Guy H. Orcutt (1980)
James N. Morgan (1981)
Donald Brown (1982)
Richard Freeman (1982)
Glenn Loury (1982)
Vincent McDonald (1982)

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION
Elizabeth Bailey, *Chair* (1982)
Marianne Ferber (1980)
Barbara A. Jones (1981)
Helen F. Ladd (1981)
M. Louise Curley (1982)
Robert Eisner (1982)
Nancy Ruggles (1982)
Moses Abramovitz, *ex officio* (1980)

COMMITTEE ON U.S.–CHINA EXCHANGES
Dwight H. Perkins, *Chair* (1981)
Hollis Chenery (1981)
Gregory Chow (1981)
Robert Dernberger (1981)
John G. Gurley (1981)
Lawrence R. Klein (1981)
Tjalling C. Koopmans (1981)
Robert Solow (1981)
Benjamin Ward (1981)

COMMITTEE ON U.S.–SOVIET EXCHANGES
Lloyd G. Reynolds, *Chair* (1982)
Abram Bergson (1982)
Joseph Pechman (1982)
Richard Rosett (1982)
Rendigs Fels, *ex officio*

Council and Other Representatives

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE SECTION K ON SOCIAL AND ECONOMIC SCIENCES
Roger Bolton (1982)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SLAVIC STUDIES
Elizabeth Clayton (1982)

AMERICAN COUNCIL OF LEARNED SOCIETIES
C. Elton Hinshaw (1981)

FEDERAL STATISTICS USERS CONFERENCE
Paul Wonnacott (1982)

INTERNATIONAL ECONOMIC ASSOCIATION
Anne O. Krueger (1984)
C. Elton Hinshaw (1985)

INTERSOCIETY COMMITTEE ON TRANSPORTATION
William Dodge

POLICY BOARD OF THE JOURNAL OF CONSUMER RESEARCH
Lester Telser

NATIONAL ARCHIVES ADVISORY COUNCIL—GENERAL SERVICES ADMINISTRATION
William N. Parker (1981)

NATIONAL BUREAU OF ECONOMIC RESEARCH
Carl F. Christ (1981)

SIXTH SYMPOSIUM ON STATISTICS AND THE ENVIRONMENT — STEERING COMMITTEE
Eugene Seskin

SOCIAL SCIENCE RESEARCH COUNCIL
Finis Welch (1981)

SSRC – COMMITTEE OF PROFESSIONAL ASSOCIATIONS ON FEDERAL STATISTICS (COPAFS)
John H. Cumberland (1980)
Gary Fromm (1981)

Representatives of the Association on Various Occasions—1980

INAUGURATIONS

D. Bruce Johnstone, State University College at Buffalo
Blair C. Currie

Ronald K. Calgaard, Trinity University
Claude A. Talley, Jr.

Gloria McDermith Shatto, Berry College and Berry Academy
Siegfried G. Karsten

Raleigh Kirby Godsey, Mercer University
JoAnn Jones

Harold Tafler Shapiro, University of Michigan
David J. Smyth

Lauro Fred Vavazos, Texas Tech University and Health Sciences Center
Barry L. Duman

Donald W. Zacharias, Western Kentucky University
Charles W. Campbell

Ralph M. Tanner, Baker University
Morris L. Stevens

Richard Earl Berendzen, The American University
Bradley B. Billings

James Gordon Kingsley, William Jewell College
Paul H. Engelmann

Arnold R. Weber, University of Colorado
Doris M. Drury

Oscar E. Remick, Alma College
John P. Henderson

ASSA 1980 Convention Committee

Thomas E. Davis, *Chair*
Glenn Miller, *Vice-Chair*
Barbara Weaver, Convention Manager
Tucker H. Adams
Norma J. Ayres
Cathie Collins
Kathleen M. Cooper
Doris M. Drury
Scott Hoober
Evan Laman
Susan F. Norman
Barry K. Robinson
Violet O. Sikes
Melissa A. Williams
Mary L. Winer

C. ELTON HINSHAW, *Secretary*

Report of the Committee on Elections

In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee, the Electoral College, and the Committee on Elections.

The Nominating Committee, consisting of Lawrence Klein, Chair, Herman E. Daly, David A. Kendrick, Sherwin Rosen, Anna J. Schwartz, Lester Thurow and Murray L. Weidenbaum, submitted the nominations listed below for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect
Gardner Ackley

Vice-Presidents	*Executive Committee*
Otto Eckstein	Elizabeth E. Bailey
Leonid Hurwicz	Robert J. Gordon
Dale W. Jorgenson	Jacob Mincer
Alice M. Rivlin	Sherwin Rosen

The Secretary prepared biographical sketches of the candidates and distributed ballots in late summer. The Committee on Elections, consisting of Ben W. Bolch, Chair, and C. Elton Hinshaw, *ex officio*, canvassed the ballots and filed the following results:

Number of envelopes without names for identification 243
Number of envelopes received too late 16
Number of legal ballots 5,483
5,742

On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)
Gardner Ackley

Vice-Presidents (for a term of one year)
Otto Eckstein
Alice M. Rivlin

Members of the Executive Committee (for a term of three years)
Elizabeth E. Bailey
Robert J. Gordon

In accordance with the actions of the Executive Committee at its meeting on December 27, 1979, an amendment to Article I, Section 2 of the bylaws was submitted to the members in a mail ballot in conjunction with the balloting for officers. The ballots were canvassed by the Committee on Elections. On the basis of the canvass, I certify that the amendment was approved.

The bylaws as amended now read:

Article I, Section 2.

Effective January 1, 1976, the base fee is $25.00 per year. The Executive Committee may increase the dues schedule, including both the base fee and the income brackets for regular members, in proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes.

BEN W. BOLCH, *Chair*

Report of the Treasurer for the Year Ending December 31, 1980

Thanks to a decision by the Executive Committee on September 4, 1980, the financial position of the American Economic Association continues to be strong. As a result, dues increases in the near future will continue to be held below the general rate of inflation. The Executive Committee authorized me to raise the price of annual subscriptions to the *American Economic Review* and the *Journal of Economic Literature* to $100 effective January 1, 1981, if after investigation I concluded that such an increase was warranted. I found that $100 would be low compared to similar journals in the natural sciences and economics. At the time, the *Journal of Inorganic and Nuclear Chemistry* cost $440, the *Journal of Biological Chemistry* $260, and the *Journal of Econometrics* $242. Those may be extreme cases, but for 1981, the sum of the subscription rates for the *Quarterly Journal of Economics* and the *Review of Economics and Statistics* will exceed the proposed $100 for the *AER* and *JEL*. Believing that the demand is inelastic, I put the $100 price into effect. We expect to lose a significant number of subscribers but anticipate a large increase in revenues.

The new rate for subscriptions does not apply to members dues, the base rate for which went up from $28.75 to $30.00 on January 1, 1981. The present dues structure with higher rates for associate and full professors was put into effect on January 1, 1976, with a base rate of $25. The 20 percent increase in a period of five years is far less than the rise in the consumer price index or the *GNP* deflator.

At this writing (January 9, 1981), the financial results for 1980 are not known. The Auditor's Report and accompanying financial statements for 1980 will be published in the June 1981 issue of the *Review*. Unaudited results for the first nine months show a surplus.

A preliminary budget for 1981 was reported to the Annual Meeting of the members held on September 6, 1980. It was virtually in balance. It will be revised extensively and submitted to the Executive Committee for approval at its March 1981 meeting. There will be increases in both expected revenues (to reflect the increase in subscription rates) and costs.

RENDIGS FELS, *Treasurer*

Report of the Finance Committee*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 31, 1980, with costs and market values as of that date. The total market value of the securities portfolio at year end was $1,956,493. After making adjustments for cash additions and withdrawals, we estimate that the Association's investment portfolio experienced a total investment return of +31 percent during 1980.

At its annual meeting in late 1979, the Finance Committee reaffirmed its investment policy of establishing an equity ratio range of 50 to 75 percent. In addition, the investment manager was authorized to lengthen fixed income maturities at his discretion, but not to exceed an eight-year maturity on average.

In view of both this investment policy and the market environment, several portfolio changes were made last year. These included new commitments, still held at year end, in Litton Industries, Humana, SmithKline, Central Louisiana Energy, Crown Zellerbach, Honeywell, Texas Instruments, Warner Communications, and SCA Services, and the elimination of Central and Southwest, Corning Glass, Northern Telecommunications, Minnesota Mining, McDonalds, Continental Illinois, John Deere, Ft. Howard Paper, and First Bank System. These portfolio changes and the year's market appreciation resulted in the portfolio's equity ratio being near the upper end of its permitted range. It actually closed the year at 76 percent. Maturities in fixed income securities were increased modestly and presently have an average effective length of about 4.4 years.

In terms of the portfolio's investment experience, the Committee can also report that, in addition to the full portfolio's return of +31.6 percent for the year, the Association's equities taken alone had a total return of +40.1 percent. This was substantially greater than that of the widely followed market averages. (Total return for the Dow Jones Industrial Average and the Standard and Poors *500* was +22.4 percent and +32.6 percent, respectively.)

In terms of future investment policy, the Finance Committee decided at its meeting in December to continue the 50–75 percent equity operating range but to allow the Association's investment counsel to use its discretion as to whether to sell equities when a rising stock market causes the portfolio to exceed the upper limit.

ROBERT EISNER, *Chair*

*The report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which will appear in the June 1981 issue of the *Review*.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1980

	Value	Percent	Estimated Income	Estimated Current Yield
Cash Equivalents	188,623	9.6	29,038	15.4
Short-Term Securities	92,626	4.7	9,375	10.1
Medium-Term Securities	141,625	7.2	16,950	12.0
Long-Term Securities and Preferred Stocks	46,600	2.4	5,625	12.1
Convertible Securities	0	0.0	0	0.0
Equity Securities	1,487,019	76.0	35,771	2.4
Total	1,956,493	100.0	96,759	5.0

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1980

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
Cash Equivalents and Fixed-Income Securities (24.0 percent)						
CASH EQUIVALENTS (0–1 year) (40.2 percent)						
Cash			31,583		31,583	5,259
Stein Roe Cash Reserves, Inc.	118,139	1	118,140	1	118,140[a]	19,729
U.S. Treasury Notes (10.125 09/30/81)	40,000	97	38,900	100	39,940	4,050
	158,139		157,040		158,080	23,779
Subtotal Cash Equivalents (0–1 Year)			188,623		189,663	29,038
Other Short-Term Securities (1–5 years) (19.7 percent)						
Fed. Nat. Mtg Assn (9.500 03/10/83)	50,000	94	46,813	98	49,141[a]	4,750
U.S. Treasury Notes (9.250 05/15/84)	50,000	92	45,813	96	48,094	4,625
	100,000		92,626		97,235	9,375
Subtotal Other Short-Term Securities (1–5 years)			92,626		97,235	9,375
Medium-Term Securities (5–10 years) (30.2 percent)						
Fed. Farm Cr. Banks (10.750 10/20/86)	50,000	92	45,750	100	49,859	5,375
Fed. Nat. Mtg Assn (11.150 05/11/87)	50,000	93	46,500	90	44,781	5,575
U.S. Treasury Notes (12.000 05/15/87)	50,000	99	49,375	100	49,871	6,000
	150,000		141,625		144,511	16,950
Subtotal Medium-Term Securities (5–10 years)			141,625		144,511	16,950
Long-Term Securities (More than 10 years) (9.9 percent)						
Hydro-Quebec Debentures (11.250 10/15/09)	50,000	93	46,600	96	48,031[a]	5,625
Subtotal Long-Term Securities	50,000		46,600		48,031	5,625
Total Cash and Fixed-Income Securities			469,474		479,440	60,988
Equity Securities (76.0 percent)						
Extractive-Energy (1.4 percent)						
Mapco	500	43	21,250	18	8,855	850
Energy Services (18.4 percent)						
Halliburton	1,000	84	83,500	32	31,897[a]	1,200
Ocean Drilling and Exploration	3,800	50	190,000	10	39,699[a]	3,040
			273,500		71,596	4,240
Food, Beverages and Tobacco (2.3 percent)						
Philip Morris	800	43	34,600	22	17,726	1,280
Paper (3.3 percent)						
Crown Zellerbach	1,000	49	48,500	46	46,228[a]	2,300
Drugs and Hospital Supplies (9.3 percent)						
Abbott Lab	1,000	57	56,500	21	21,360[a]	1,200
Merck	500	85	42,375	57	28,402[a]	1,300
SmithKline	500	80	40,000	57	28,333	960
			138,875		78,095	3,460
Petroleum (15.8 percent)						
Atlantic Richfield	1,000	64	63,625	36	35,823	1,900
Cities Service	1,000	48	47,750	17	17,185[a]	1,600
Conoco Inc	800	65	52,300	19	15,580[a]	1,760
Gulf Oil	800	44	34,800	17	13,321	2,000
Standard Oil Ohio	510	72	36,720	20	10,076[a]	918
			235,195		91,985	8,178
Computers and Office Equipment (4.1 percent)						
Honeywell	250	112	27,938	91	22,833	750
IBM	480	68	32,581	28	13,325[a]	1,651
			60,519		36,158	2,401
Electrical Equipment (2.8 percent)						
General Electric	690	61	42,263	36	24,536[a]	2,070
Electronics (1.6 percent)						
Texas Instruments	200	121	24,150	118	23,616	400

TABLE 2–(*Continued*)

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
Photography (2.4 percent)						
Eastman Kodak	500	70	34,875	47	23,740	1,750
Broadcasting (1.6 percent)						
CBS	500	48	23,813	37	18,662[a]	1,400
Natural Gas Companies (4.1 percent)						
Central LA Energy Corporation	1,200	51	61,200	31	36,750	1,200
Sanitary Services (1.7 percent)						
SCA Services Inc	1,500	17	25,125	18	27,369[a]	
Other Financial Services (2.3 percent)						
Alexander and Alexander	1,000	34	34,000	9	9,325[a]	1,640
Services (9.9 percent)						
Disney	700	51	35,875	22	15,503[a]	700
Humana Inc	900	71	64,238	37	33,677[a]	1,080
Warner Communications	600	78	46,575	49	29,223	816
			146,688		78,403	2,596
Conglomerates (3.1 percent)						
Litton Industries	510	89	45,326	53	27,115	612
Miscellaneous (16.0 percent)						
Stein Roe Capital Opportunities Fund	2,020	26	52,626	20	40,000	505
Stein Roe Special Fund	3,292	15	48,230	12	40,000	889
Stein Roe Universe Fund Inc	2,466	55	136,284	51	125,000[a]	
			237,140		205,000	1,394
TOTAL EQUITY SECURITIES			1,487,019		825,159	35,771
TOTAL SECURITIES AND CASH			1,956,493		1,304,599	96,759

[a] More than one cost basis.

Report of the Managing Editor

American Economic Review

This is my final report as managing editor of the *Review*. While most of my remarks will be devoted to matters of editorial policy, I wish to summarize briefly the operations during 1980.

The level of submissions fell this year. Our office listed 529 papers received (see Table 1). In addition 112 papers were received in the West Coast office, after it began its operations October 1, 1980. We published 127 papers, more than the last few years, and ran slightly larger issues than last year, 1137 pages over 1058 (see Table 2). The purpose of the larger 1980 issue was to limit the backlog of accepted papers that Robert Clower would have to publish before seeing his own acceptances in print.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1961-80

Year	Submitted	Published	Ratio of Published to Submitted
1961	305	47	.15
1962	273	46	.17
1963	329	46	.14
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	529 112} 641	127	.20

It is too soon to tell whether the smaller number of submissions in 1980 signifies a permanent decline in volume. As long as the number of publishable papers remains intact no one will complain if the volume is reduced.

As shown in Table 3, the distribution of submitted and published papers remains virtually unchanged from prior years. The most popular fields continue to be microeconomics, welfare theory, international economics, macro and monetary economics, and labor.

I. Board of Editors

The Board of Editors consists of eighteen members, chosen by the managing editor, with the approval of the executive committee of the Association. Their names are printed on the contents page in every issue. The Board has been responsible for two of the most parts of the refereeing process: determining the quality of comments on published articles; and reading papers that are the subject of complaint over the fairness of referees. The cooperation of the Board has been excellent.

TABLE 2—SUMMARY OF CONTENTS, 1979 AND 1980

	1979		1980	
	Number	Pages	Number	Pages
Articles	52	635	52	680
Shorter Papers, including Notes, Comments and Replies	67	353	75	392
Dissertations		26		18
Announcements and Notes		35		36
Index		9		11
TOTAL	119	1058	127	1137

TABLE 3—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS IN 1980

	Submitted	Published
General Economics and General Equilibrium Theory	13	2
Micro-Economic Theory	90	21
Macro-Economic Theory	46	14
Welfare Theory and Social Choice	42	19
Economic History, History of Thought, Methodology	10	1
Economic Systems	6	1
Economic Growth, Development, Planning, Fluctuations	22	8
Economic Statistics and Quantitative Methods	21	8
Monetary and Financial Theory and Institutions	37	9
Fiscal Policy and Public Finance	21	5
International Economics	80	14
Administration, Business Finance	20	1
Industrial Organization	27	4
Agriculture, Natural Resources	14	0
Manpower, Labor Population	57	12
Welfare Programs, Consumer Economics, Urban and Regional Economics	23	8
TOTAL	529	127

The Board also advises the managing editor on matters of editorial policy, techniques of editorial control, and the content of articles. They have been very useful, and I am grateful to them for their hard work and warm interest in the *Review*.

In March 1980, five new members of the Board were appointed by the Executive Committee for three-year terms. They are Herschel Grossman, Peter Howitt, Ann Krueger, James Smith, and Robert Willig. Albert Ando has agreed to accept a second term.

Six members of the Board will complete their terms at the end of 1980: Rudiger Dornbusch, William Oakland, Richard Roll, Michael Spence, William Vickrey, and S. Y. Wu. I wish to thank them for their high professional standards, work, and cooperation. I also wish to acknowledge with thanks the services of the continuing members: Pranab Bardhan, Peter Diamond, W. Erwin Diewert, Michael Parkin, Roy Radner, and Nancy Schwartz.

II. Screeners, Referees, and Proofreaders

I wish to thank the following graduate students who have worked this year as proofreaders and hunters of false proofs: George Briden, Inhak Lim, Kee Park, and Joel Scheraga.

My thanks also to the following economists who have served as editorial consultants in the screening of manuscripts: James Albrecht, Theodore Bergstrom, Roger Bolton, George Borjas, Anthony Cassese, Arthur Denzau, Gary Dorman, Benjamin Eden, Roger Feldman, Allan Feldman, Donald Frey, H. Landis Gabel, John Geweke, Gerald Goldstein, Edward Green, Donald Hanson, Milton Harris, Robert Hodrick, Charles Lieberman, Lucas Papademos, Owen Phillips, Raymond Riezman, John Roberts, Harvey Rosen, Thomas Russell, John Rutledge, Andrew Schotter, Steven

TABLE 4—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING: 1980 *AER*

	Copies	Pages		Cost		
Issues	Printed	Net	Gross	Issue	Reprints	Total
March	28,000	268	320	$44,374.92	$1,479.14	$45,854.06
May	27,500	489	520	65,774.57	2,397.46	68,172.03
June	27,500	291	304	40,012.31	1,636.31	41,648.62
September	27,500	300	336	46,997.69	1,704.99	48,702.68
December	27,500	298	352	48,448.51	1,949.77	50,398.28
Annual Misc.[a]						3,161.04
TOTAL		1,646	1,832	$245,608.00	$9,167.67	$257,936.71

[a] Includes cost of preparing mailing list, extra shipping charges, and storage costs of back issues.

Steven Shavell, Robert Shishko, Charles Stone, Allyn Strickland, John Trapani, Bernard Wasow, Louis Wilde, Kenneth Wolpin, and Allan Zelenitz.

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of economists during the year. I wish to thank them for their cooperation and high standards in reading and evaluating manuscripts. Their names are listed at the end of the full report.

III. Editorial Policies and Management

I wish to use the occasion of my final report to portray the policies and standards that served as my guides over the last twelve years. My predecessors left very little behind to help the historian of economic thought or the professional gossip. There is little in print to answer the obvious questions one might put to the editor: What is the role of the *Review*; what is the editorial policy; what is its contribution to economic knowledge. We do know that each editor, on putting down the blue pencil, issued a sigh of relief. Paul Homan must have understated his feelings in his last report: "...There are pits into which the editorial judgment may fall. I count it some sort of triumph, more of the human spirit than of my performance, that I have been denounced only twice to this Committee and can count serious controversies on the fingers of one hand" (*AER*, May 1952, p. 744).

John Gurley's farewell was simple, "Let George do it." Mine will be thirteen pages longer and a bit more contentious. This report will cover three related subjects: the management of the *Review*; its content and its influence on economics; and the conflicts in editorial policy.

By and large, the *Review* is limited to print what is submitted. There are a few papers commissioned by the editor each year; the majority are selected through a process of screening and refereeing. When a manuscript is received, it is sent out to be screened to determine if it is of potential interest. The screener is asked to identify those papers that are unoriginal, of a low level of intellectual interest, or are perhaps more appropriate for other journals. The screener also writes a brief summary of each paper. A fee of $15 per manuscript is paid to the screener.

The purpose of screening is to reduce the fraction of the editor's time spent on poor quality papers that will not be published. My predecessor, John Gurley, complained about this aspect of editing, and before I introduced screening, my experience confirmed this. Last year I used the services of forty-seven economists as screeners. They were selected by recommendation of the Board of Editors or department chairmen. The screeners are younger members of the profession who have completed their Ph.D.s, perhaps published an article, and are willing to work for a very small fee. A few of the screeners have stayed on for a number of years, and have developed considerable editorial skill.

After the papers are returned by the screener, I generally send out to referee the 75 percent so recommended, and try to find refereeable papers among the 25 percent the screeners have suggested are not appropriate. Some of the papers rejected at the screening stage will be resubmitted after revision.

Refereeing is a more detailed process than screening. The referee is asked to work through the paper to determine its quality, originality, and suitability for publication. The referee is asked to write a report explaining the basis for the decision, providing suggestions to the author. The referees are not paid. They do this work to keep up in their field and as a response to the challenge of new ideas.

Neither the screener nor the referee need know the identity of the author. All papers are sent out with the author's name removed from the title page and the biographical reference.

Most of the refereeing is carried out by individual scholars. Specialized refereeing is done by the Board of Editors. In my 1979 report, I listed the names of 388 individual referees. In 1979, 719 papers were received; 180 were rejected after screening; the remaining 539 were sent to referee; 119 were published.

The Board of Editors currently consists of eighteen economists each serving a term of three years. They carry out special types of refereeing: they evaluate comments on published work, papers that are the subject of complaint by the author, and papers that for some reason cannot find a patient referee. The functions of the Board have evolved over time. Under earlier editors, and earlier in my term, they were the major refereeing resource of the *Review*. However, I have found the Board to be more valuable in the above-described capacity, where maturity and thoughtful advice can be enormously helpful. Members of the Board of Editors are selected from among the most productive scholars in the profession. At the end of this report, I list all of the Board members with whom I have served. I also list the referees who have evaluated over fifteen manuscripts.

The editor plays a role in the selection of manuscripts for publication. More papers are recommended favorably by referees than the *Review* can print. In addition, referees will reject for reasons that are not articulated, or are subject to disagreement. In the first case, I try to weed out and reject otherwise acceptable papers in areas that have received thorough exposure. The authors are told the paper is publishable and advised to send it elsewhere. They do get published elsewhere. I have disagreed with a referee's negative evaluation, when I saw something in the paper or in the screener's report that suggested an interesting idea. It may also be the case that the referee has given the paper superficial treatment and has not explained satisfactorily the reason for rejection. I will then send the paper out to be read again.

An acceptable paper should represent a contribution to knowledge. It should present a new way of looking at some aspect of economic analysis, institutions, data, or policy; it should present new hypotheses, or new ways of thinking about old hypotheses. It can be an expository paper, or it could bring together and synthesize ideas which seem to be unrelated.

In evaluating papers, I don't consider them as falling into methodological schools, for example, neoclassical or radical; monetarist or Keynesian. I do have a bias in favor of papers that contain some underlying rational explanation of the phenomena being analyzed; rational meaning consistent with a set of assumptions about intelligent behavior.

The most aggravating papers to judge are the comments and replies. When comments are accepted on a previously published paper, the author is invited to reply. There is little editorial control over the content of replies, and the privilege is frequently abused. Some authors write churlish replies, or they try to cover their tracks. Others are graceful and leave the reader with a good feeling. The comments require special refereeing, since the referee must be familiar with the original paper and with the field itself. Thus the Board of Editors is a happy choice to referee submitted comments. Writing a comment is an art form of its own. Some authors approach the task as if it were an execution; others as if it were a duel. I advised one author to use an epee, not a hatchet.

The publication of comments forces the editor to decide who should be invited to reply, or whether a reply is called for at all. Sometimes a comment is addressed to the work of a number of authors; at other times the comment is addressed to a framework of analysis that should be considered in the public domain. When the editor chooses not to invite a reply or chooses to invite *A* and not *B*, he will receive interesting mail. He will also receive unsolicited advice and replies from outraged authors who claim the proprietary right of reply when their work is mentioned with insufficient praise.

IV. Content and Influence

The content of the *Review* has changed over time. To see this evolution, one should go back to the beginning when Davis Dewey started as the first managing editor in 1911. At the time of his retirement in 1940, Dewey described his policy and noted his reluctance to take on the job:

> When I was invited to take the managing editorship,... I demurred on the ground that my chief interest was in American economic problems and

> not in the refinements of economic theory. My acquaintance with theory was limited to some knowledge of Adam Smith, John Stuart Mill, Karl Marx, and Francis A. Walker. I had tried to keep pace with the newer Marshallian analysis and to reconcile the reasoning of the Austrian school with the antiquated concepts of pre- and mid-Victorian economists. And such economics as I had imbibed was imbedded in a thick layer of Vermont GOP. Thus you can see that I was but poorly qualified to assume the editorial responsibilities which the year 1910 demanded. My shortcomings in theory were met by the answer that the *Quarterly Journal of Economics* ably took care of theory.
>
> [*AER*, Feb. 1941, p. *viii*]

Under Dewey's editorship, the *Review* had a strong practical and institutional flavor. It published papers in taxation, public utility regulation, wage policy and labor markets, business fluctuations, financial markets, competition and monopoly, banking, and tariffs. However the great articles of the 1930's and 1940's on ordinal utility, indifference curves and demand functions, on cost curves and supply curves, on tariffs and welfare, on savings, investment, and macroeconomic equilibrium, did not appear in the *Review*. The qualitative impression that I took with me out of graduate school was that the major contributions in economics appeared elsewhere. I had the occasion to confirm this impression by tabulating the origin of articles included in books of reprints. I selected the ten books of *Readings* that appeared under the aegis of the Association between 1942 and 1969. In two of these volumes, *Readings in Fiscal Policy* (1955) and *Readings in Industrial Organization and Public Policy* (1958), articles from the *Review* were dominant: 15/34 in the first and 13/21 in the second. After that it goes downhill: *Readings in Welfare Economics* (1969) had 0/39; *Readings in The Theory of International Trade* (1949) had 1/23. The others are stretched out in between:

Readings in		
Theory of Income Distribution	(1946)	5/32
Social Control of Industry	(1942)	3/15
International Economics	(1968)	3/33
Price Theory	(1952)	3/25
Monetary Theory	(1951)	5/20
Business Cycle Theory	(1944)	4/21

I also checked the bibliographies of a number of well-known economists. The first three volumes of Paul Samuelson's *Collected Papers* contain 204 papers, of which 18 appeared in the *Review*. Don Patinkin's article on Frank Knight (Dec. 1973) cites 27 major works; one appeared in the *Review*. Similar results were obtained from an examination of bibliographies that accompanied survey articles in the 1960's: in E. J. Mishan's survey of Welfare Economics (*Econ. J.*, 1960), the *Review* had 19/250; in Harry Johnson's survey on Monetary Theory (*AER* 1962), the *Review* had 12/130.

The above citations for the most part cover articles published prior to 1960. I think there has been a dramatic change in the content of the *Review*, and it started after World War II. Papers in theoretical topics have increased in number, although mathematical complexity increased more slowly. More recently there has been a flowering of theoretical articles in the *Review*, and it has carried original work in diverse fields, including welfare theory, international trade and finance, monetary theory, business finance, regulation, theory of the firm, consumer behavior, economic growth and development, optimal taxation, and choice under uncertainty.

The question of what the *Review* should be is never settled. Prior to 1969, the *Review* was the Association's only major publication, and carried articles, notes, book reviews, classified lists of new books, periodicals, and dissertations. In the early 1960's, a new journal was financed by the Association. It was called the *Journal of Economic Abstracts*, and edited by Arthur Smithies. In 1969, the *JEA* ceased publication, and the *Review* itself split into two parts. One part became the *Journal of Economic Literature*. The new-born *AER* now contains articles, notes, and the Ph.D. dissertation list. One purpose of the split was to provide more space for journal articles. Before the split, the *Review* published 93 articles, notes, and communications in 1968. After the split in 1969, it published 121.

When I took over the *Review* in 1969, I had a brief period of time for reflection, because the first three issues had been reserved for the backlog of papers accepted by John Gurley. I noted that most of the *Review's* articles consisted of refereed papers, and that there seemed to be a gap. There were few articles on policy issues and few expository papers. This was hardly surprising, since economists are notoriously reluctant to invest time in writing serious policy papers that will be submitted for refereeing and subject to outright rejection. For one thing, the delays required for refereeing can reduce the timeliness of a policy paper.

I have on occasion tried to remedy this by commissioning articles on issues of economic policy. The first articles were reviews of the annual report of the President's Council of Economic Advisors. These were well done and interesting, but faded in the mid-1970's with the decline of the analytic content of the Report. I then solicited review articles on policy reports by the Federal Reserve Board and the Council of Environmental Quality. A second type of policy paper was suggested by the opportunity to observe what was at the time an unusual event. I invited four English economists to write on their perceptions of the causes of the British inflation of the mid-1970's. These were very successful papers and they served me well as teaching material.

I also thought it useful to invite papers on the history of economic thought, and have invited all the American recipients of the Nobel award to publish their acceptance addresses in the *Review*. Indeed we have carried the addresses by Paul Samuelson, Simon Kuznets, Kenneth Arrow, Wassily Leontief, Tjalling Koopmans, Herbert Simon, and W. Arthur Lewis.

Still another type of invited paper that is useful for the history of economic thought is the memorial article or memorial notice. I have printed memorial notices (two to three pages) for deceased presidents of the Association, and printed a rather full memorial article by Don Patinkin on Frank Knight as a teacher of economics. There is one deceased president who remains unshriven as I search among his former colleagues for a willing memorializer.

V. Issues of Editorial Policy

A

The *Review* is one of a large number of scholarly journals in economics, published in the United States and abroad. It has the distinction of the widest circulation (shared with *Journal of Economic Literature*) because it is sent to all members of our Association. There are differing views on the menu our readers should be offered, and these differences go back to the founding of the journal. The policy that I have followed, and I believe inherited from my immediate predecessors, is to seek and to accept the highest quality of scholarly research in any field of economics. It is my belief that the *Review* should be the flagship journal of the profession, and should publish the best that we have to offer in the way of original research. This places us in competition for manuscripts with almost every other scholarly journal in the profession, because we publish in all of the fields in which high quality work is going on. It means that readers of specialized journals (for example, in such fields as labor, money and banking, international economics, etc.) will also find publishable material in the *Review*, and that authors in these fields will have both specialized and general outlets for their work.

A general journal of scholarly research serves the economics profession in a number of important ways: First, it communicates theoretical and conceptual developments across specialties so that economists doing research in one specialty may remain up to date on developments in other specialties, but more important, may carry into their own work the fruitful ideas and approaches of others. This cross fertilization of fields accounts for the rash of articles that will sometimes appear, applying a hypothesis derived in one area to work in another. Second, the general journals prevent the subspecialties from being dominated by any particular clique, in-group, or band of friends espousing a particular point of view, or methodology. Third, a general journal gives wide circulation to original research, as compared to the readership of specialized journals, thus providing psychological and material benefits to their authors. The importance of this benefit cannot be underestimated. Most of our authors are younger members of the profession, and most of the papers we print are distillations from doctoral theses. The older, well-established authors have many opportunities to publish outside the realm of the highly competitive refereed journals. Fourth, a general journal can keep its readers in their role as teachers informed and up to date on the state of thought in the various specialties of economics. If an economics teacher did not read the journals, the quality of his lecturing would have to suffer over time, no matter what student grade level was being taught.

Granting the professional contributions of the general journal of scholarly research, why should the American Economic Association run such a journal? The question has arisen more than once in deliberations over the responsibilities of the Association to its members. There are alternative menus that might be offered to the membership. One is a journal devoted to expository papers, policy issues, and teaching materials. Under this alternative, research at the frontiers of the discipline would be the purview of the specialized journals and the other general journals. The Association would confine itself to publications that had the widest readership appeal. Presumably, *JEL* would remain intact under this view since it consists of book reviews, commissioned expository and review articles, and annotations and abstracts. Each of these attract a large audience, while a scholarly research paper will reach a smaller fraction of readers.

When this alternative has been suggested, I have strenuously opposed it. I think the chief function of the Association (the promotion and dissemination of economic ideas) is best performed through the stimulation and dissemination of economic research. While individual research papers may each attract a small audience, their cumulative effect must be to communicate changes in the way economists define and approach their subject. Moreover, the alternative menu is a true alternative only if the *Review* were to refuse to publish original research. There is no reason why the expository and policy material cannot appear in the same journal that publishes research.

A final reason in favor of an Association journal is its independence. Its Board of Editors and its referees are drawn from universities and research institutions all over the United States and Canada. It is not the outlet of any one school, university, or group.

B

One of the most notable developments in our discipline is the spread of economic theory through the use of formal mathematical models. Indeed, one need only look back at the reviews of Samuelson's *Foundations of Economic Analysis* and at the mathematical content of the journals printing those reviews in 1948 to realize how much has changed. Today's journals are the carriers of this change, and the differences among journals are far narrower than in 1948.

The increased use of mathematics has made life more difficult for the editor and the referees. There was a time when a paper heavy in math content would be refused by most journals and published if at all by only a small number. Such papers were looked on by most editors as unintelligible, and they did not choose to face the problem of evaluating quality. This posture is no longer possible. Math is a language like English, and papers conveyed in math must be worked through. But the editor must keep in mind that a paper heavy in math may be just as confused as a paper with no math. Moreover if the paper is accepted, the editor must decide what parts are to be published, and whether the math stays in the text or goes into an appendix. A few years ago, on accepting a paper, I suggested that the author reduce its size by eliminating all mathematical proofs, offering to provide them to interested readers. The paper was more readable, but a critical subscriber complained that I was turning the *Review* into a journal of unproved conjectures.

I have on occasion turned away papers because their contribution seemed exclusively mathematical, and contained little of interest to general economists. Nonetheless readers will complain that our papers contain too much math. They may be correct, but for a reason they don't suspect. The first time an idea is worked out and presented, it may be far less elegant and may contain a far larger number of mathematical steps than will be required after it is refined at the hands of several authors. Math content is the price paid for publishing original work and the work of younger authors.

The spread of mathematics is a part of a more general problem faced by the editor; namely, how to provide an appropriate mix of the various research methodologies and research problems. I think of three broad categories of research problems: theory; policy; and empirical investigations. I have already indicated that economists appear reluctant to invest time in policy papers, although they are quite willing to point out the policy implications of whatever they are doing. My own preference in choosing among theory and empirical papers has been to give preference to papers that emphasize the rational, choice theoretic aspects of whatever behavior is under investigation. This is in keeping with the type of research that is being done in the profession. Few authors send in empirical investigations that are unaccompanied by a theoretical framework.

C

One of the most vexing complaints leveled at the *Review* is that it serves a neoclassical club and that the editor discourages by example submission of research based on other methodologies. To be quite specific, the complaint has been levied that the *Review* excludes the work of Marxist, radical, and post-Keynesian authors. While its pages are open to the works of neoclassical economists, they are closed to the followers of other paradigms. What is the truth? If you look at the contents of the *Review* over the last twelve years (as well as earlier), you will see a notable absence of papers by Marxist, radical, and post-Keynesian authors. The symptoms are correctly noted. What is the pathology? I can see two possible explanations: first, that Marxist, radical, and post-Keynesian authors do not wish to publish in the *Review* and therefore do not submit their work; second, that the papers of these authors have been submitted and have been systematically rejected. On the first point, it is true that very few MRPK papers are submitted. It is not possible to keep a number count of the papers classified by research methodology, or paradigm. My memory of authors' names, however, leads me to conclude that the *Review* is not a popular place for the authors of such papers. I have indeed published one paper on Marxian economics and did publish a radical critique of one of the Council of Economic Advisors' reports. Neither paper attracted any publishable comments or any great attention. The second explanation is also partially true; whatever papers I have received from MRPK authors have been rejected by the refereeing process. Again we are talking of a very small number of papers. What is the truth? Is there a plot to exclude such papers, or is it the case that there simply is not much good work going on in these three areas? It is appropriate to raise the parallel with the difficulties experienced some time ago by authors of mathematically oriented papers. They founded their own journals, and their methodology has now captured the profession. Economics is a highly competitive field, and if any methodology has powerful medicine, we are sufficiently opportunistic to wish to use it in our research. My advice to MRPK authors is to compete.

Nevertheless, there is a challenge to the *Review*:

a) How to keep the minds of the editor, the Board of Editors, and referees open to work that is not in the neoclassical tradition. A corollary is how to recognize truly original work if it has no neoclassical origins. The problem is to find good original work. We are not turning away good papers by radicals. We are not getting any.

b) The second problem is how to convince readers and potential authors that there are no editorial biasses in the selection process, particularly when the *Review* contains so few articles in radical economics. One attempt in this direction was made in the September 1972 issue of the *Review*, in a statement of editorial policy which was reprinted in the May 1973 *Proceedings*. That statement did not satisfy any critic. How can a thousand flowers bloom if they are not all watered?

c) A third problem is that the substantial majority of potential authors and referees are not working in the radical tradition and strongly question the wisdom of allocating space to an effort which they feel has a strong likelihood of proving unfruitful.

The radical political economists, post-Keynesians, and Marxists have their own journals. If their methodologies are fruitful, I would expect economists of all persuasions to use them.

D

Some things are worth noting because they didn't happen. During most of the period of my editorship, the learned professions in this country have been wracked by moral debates and proposals for social action. Opposition to the war in Vietnam, demands that the professions recognize the rights of women and racial minorities came to the agenda of our annual meetings and split many of the other societies. The economics profession went through this period with less agony, because of the willingness of all parties to find middle ground. Very little of this conflict and demand for social action rubbed off on the *Review*. I can think of only one minor incident, and mention it to indicate how little strife we experienced. Early in my first term as editor I was accused by a radical author of having appointed a well-known war criminal to the Board of the Review. The editor in question had committed the offense of taking leave from his university to serve at a high rank in the federal government in the Nixon Administration.

E

A final word on editorial control. The managing editor has two primary sources of advice for editorial policy. One is the Board of Editors of the *Review*, with whom he meets annually and corresponds occasionally. The second is the Executive Committee of the Association with whom he meets twice a year. The work of the Board of Editors has been described earlier. They provide a sounding board for new ideas, complaints, and suggestions dealing with editorial policy, the mechanics of handling manuscripts, ideas for new papers, and so on. Other journals may use the Board of Editors more directly as managers of the refereeing process. I have not tried to do so. My relation with the Board has always been warm and cooperative. I think of them as a distinguished faculty, and any department chairman who reads the list of Board members appended to this report will share that feeling.

My relation with the Executive Committee is somewhat different. The managing editor serves at their pleasure, and the relation is much like that between the Board of Trustees of a prestigious college and its director of admissions. I have had excellent cooperation and support from the Executive Committees of the Association. They have acted with intelligence and restraint when complaints were brought to them, and I am grateful for their encouragement.

I also wish to thank Rendigs Fels and Elton Hinshaw for advice, wisdom, and words of calm; and thank the Presidents of the Association under whom I served. Every President has an editor's blue pencil in his knapsack, and I am grateful that so many of them refrained from attempting to use it. I didn't always agree with each of them, but then they didn't always agree with each other. Finally my warmest thanks to Wilma St. John who has endured this twelve years with fortitude and humor, and who knows more about the operations of the *Review* than any living person. I also wish to thank our office staff Debi Franklin and Sandy Overton, and wish them good luck.

In closing this report I would like to quote again from Davis Dewey. On his retirement in 1940, he left the following words of advice to his successor:

> 1) Be sure to have one article containing involved mathematical equations with unusual fonts of type. In as much as the printer has to spend a good deal of time in ransacking the type foundries of the country, this affords you a good excuse for a delay in publication.
>
> 2) Be sure that a majority of the leading articles contain at least six references to Keynes. Adam Smith, John Stuart Mill, Marshall, and Francis A. Walker and their contemporaries are now passé. And it is your duty to see that the articles you select do not burden the readers with reasoning which has been outmoded.
>
> 3) Publish at least one review in each issue which will arouse the animosity of the author. There is nothing more stimulating than controversy.
>
> 4) Be sure to have occasionally an article contain fifty-cent and one dollar words. Though difficult to understand such an article commands respect; and economists in these days need respect.
>
> [*AER*, Feb. 1941, p. *xi*]

AER Board of Editors, 1969–80

Irma Adelman
Albert Ando
Elizabeth E. Bailey
Bela Balassa
Robert Baldwin
Pranab K. Bardhan
David P. Baron
Robert J. Barro
Gary Becker
Barbara Bergmann
Jagdish N. Bhagwati
David Bradford
Phillip Cagan
Gregory C. Chow
Carl Christ
Laurits Christensen
Peter A. Diamond
W. Erwin Diewert
Rudiger Dornbusch
Eugene Fama
Martin S. Feldstein
Charles E. Ferguson
Ann F. Friedlaender
Irwin Friend

Stephen M. Goldfeld
Robert J. Gordon
Herschel I. Grossman
Robert E. Hall
Bent Hansen
Jack Hirshleifer
Peter W. Howitt
Harry G. Johnson
Ronald W. Jones
Dale W. Jorgensen
Anne O. Krueger
David Laidler
Daniel McFadden
Alvin L. Marty
James R. Melvin
Edwin S. Mills
Herbert Mohring
Marc Nerlove
William Nordhaus
G. Warren Nutter
William H. Oakland
Michael Parkin
Edmund S. Phelps

Roy Radner
Leonard A. Rapping
Stephen Resnick
Richard W. Roll
Sherwin Rosen
Thomas R. Saving
F. M. Scherer
Anna J. Schwartz
Nancy Schwartz
James P. Smith
Vernon Smith
Warren L. Smith
A. Michael Spence
Frank P. Stafford
Jerome Stein
Joseph E. Stiglitz
S. C. Tsiang
Lloyd Ullman
William S. Vickrey
Finis Welch
Marina v. N. Whitman
Robert D. Willig
S. Y. Wu

Referees Who Read Fifteen or More Manuscripts

Allen, Polly
Anderson, James
Balassa, Bela
Baldwin, Robert E.

Baron, David
Barro, Robert J.
Batra, R. N.
Baumol, William

Beckmann, Martin
Bergmann, Barbara
Branson, William H.
Brechling, Frank

Buchanan, James
Cagan, Philip
Chow, Gregory C.
Christ, Carl
Davis, Eric
Eisner, Robert
Fama, Eugene
Feldstein, Martin
Fisher, Anthony
Friedlaender, Ann
Goldfeld, Stephen
Grossman, Herschel
Hadar, Josef
Hall, Robert
Hirshleifer, Jack
Jaffee, Dwight
Jones, Ronald W.
Jorgenson, Dale
Krueger, Anne
Laidler, David
Malkiel, Burton
Marty, Alvin
Melvin, James
Meyer, Robert
Mohring, Herbert
Muth, Richard
Neher, Philip
Nerlove, Marc
Newhouse, Joseph
Oakland, William
Olsen, Edgar
Pauly, Mark V.
Phelps, Edmund S.
Poole, William
Rapping, Leonard
Resnick, Stephen
Rosen, Sherwin
Sato, Ryuzo
Saving, Thomas
Schmalensee, Richard
Schupack, Mark
Schwartz, Anna J.
Silberberg, Eugene
Smith, Vernon
Stafford, Frank
Stein, Jerome
Stiglitz, Joseph
Tsiang, S. C.
Vickrey, William
Welch, Finis
Wood, John
Wu, S.Y.
Zeckhauser, Richard

1980 Referees

P. Allen
J. Anderson
G. Archibald
F. Arditti
S. Arndt
K. Arrow
O. Ashenfelter
A. Auerbach
C. Azariadis
C. Azzi
M. Bailey
M. E. Baily
R. Baldwin
D. Baron
R. Barro
J. Barron
A. Bartel
Y. Barzel
R. N. Batra
G. Becker
W. Becker
M. Beckmann
B. Benson
G. Benston
E. Berglas
T. Bertrand
S. Bhattacharya
J. Bilson
S. Black
O. Blanchard
M. Blejer
C. F. Boonekamp
G. Borjas
M. Boskin
R. Boyer
R. Braeutigam
S. Braithwait
W. Branson
H. Brems
M. Brennan
A. Brillembourg
D. Brito
D. Bromley
M. Bruno
J. Buchanan
W. Buiter
R. Burkhauser
H. S. Burness
W. Butz
P. Cagan
G. Cain
D. Capozza
D. Carlton
H. Carter
J. Cassing
D. Caves
R. Caves
K. Chan
J. Chipman
C. Christ
L. Christensen
P. Clark
J. Conlisk
M. Connolly
R. Cooter
R. Cotterman
J. Cox
R. Craine
A. Cukierman
R. Cummings
G. Daly
S. Damus
M. Darby
R. d'Arge
S. Das
R. H. Day
R. Deacon
A. Deardorff
M. DeGroot
A. DeVany
D. Diamond
W. Dolde
M. Dooley
A. Drazen
R. Driskill
D. Dutton
C. Eaton
J. Eaton
L. Edlefson
R. Ehrenberg
I. Ehrlich
R. Eisner
B. Ellickson
J. W. Elliott
W. Ethier
E. Fama
L. Fan
H. Farber
G. Faulhaber
A. Feldman
G. Fields
S. Fischer
A. Fisher
F. Fisher
J. Flanders
B. Fleisher
R. Forsythe
R. Frank
J. Frankel
H. Frech
A. M. Freeman
R. Freeman
A. Freiden
J. Fried
B. Friedman
J. Friedman
E. Furubotn
N. Gallini
H. Genberg
A. Gifford
R. Gilbert
L. Girton
M. Goodfriend
R. Gordon

R. J. Gordon
E. Gramlich
R. Grauer
J. Gray
M. Greenhut
M. Grossman
T. Groves
J. Guasch
R. Hall
M. Hamburger
B. Hamilton
M. Hanemann
J. Hanson
W. Haraf
M. Harris
D. Hartman
R. Hartman
J. Hausman
G. Hawawini
J. Heckman
E. Helpman
W. Heller
M. Hellwig
P. Hendershott
J. Henderson
J. V. Henderson
R. Hodrick
W. Holahan
C. Holt
C. C. Holt
D. Holthausen
J. Hosek
S. Hu
D. Hueth
W. Huffman
R. Hutchens
R. Inman
M. Intriligator
D. Jaffee
L. Johnson
T. Johnson
D. Jorgenson
B. Jovanovic
G. Jump
R. Just
M. Kamien
E. Kane
B. Katz
M. Keeley
A. Kelley
R. Kihlstrom
R. King
R. Klein
P. Kleindorfer
L. Kochin
L. Kotlikoff
Y. Kotowitz
M. Kreinin
P. Krugman
D. Laidler
E. Lazear
L. Lee
H. Leland
M. Levi
K. Lewis
C. Lieberman
D. Lilien
L. Lillard
C. Lim
C. Link
R. Lucas
R. E. B. Lucas
J. McCall
B. T. McCallum
H. McCulloch
R. McKinnon
D. McNicol
L. Maccini
R. Mackay
W. Magat
B. Malkiel
A. Manne
E. Mansfield
H. Markowitz
T. Marschak
J. Marshall
S. Martin
E. Maskin
R. Masson
D. Mathieson
W. Mayer
J. Mayshar
W. Meckling
J. Medoff
L. Meyer
P. Meyer
R. Michael
M. Miles
N. Miller
F. Mishkin
B. Mitchell
H. Mohring
T. Moore
S. Morley
L. Moses
J. Muellbauer
D. Mueller
D. Mullineaux
A. Munnell
R. Muth
R. Myerson
J. P. Neary
P. Neher
M. Nerlove
J. Newhouse
P. Newman
Y. Ng
R. Noll
W. Nordhaus
W. Novshek
J. Nugent
W. Oates
R. Oaxaca
M. Obstfeld
L. Officer
M. Okuno
E. Olsen
J. Ordover
D. K. Osborne
M. Paglin
A. Panagariya
J. Panzar
R. E. Park
R. Parks
J. Paroush
D. Parsons
M. Pauly
J. Pencavel
R. Perlman
S. Perrakis
M. Perry
J. Pesando
C. Phelps
E. Phelps
L. Phlips
R. Pindyck
D. Pines
J. Pippenger
R. Pollak
W. Poole
R. Porter
A. Postlewaite
A. Protopapadakis
D. Purvis
L. Putterman
T. Rader
R. Ramachandran
G. Rausser
S. Rea
M. Reder
F. Reid
J. Richardson
J. Riley
J. Roberts
K. W. S. Roberts
R. Rohr
D. Roper
S. Rose-Ackerman
S. Rosen
M. Rosenzweig
J. Rowley
R. Ruffin
H. Ryder
L. Sahling
M. Salemi
D. Salkever
S. Salop
L. Samuelson
A. Santomero
R. Sato
T. Saving
H. Scarf
L. Schall
F. M. Scherer
B. Schiller
R. Schmalensee
A. Schmitz
A. Schotter
R. Schuler
W. Schulze
D. Schwartzman
A. Schweinberger
W. Schwert
G. Scully
J. Seater
S. Shalit
S. Shavell
R. Shiller
J. Shoven
G. Sick
J. Siegel
E. Silberberg
D. Sjoquist
D. Small
K. Smith
R. Soligo
W. Springer
D. Spulber

T. Srinivasan
F. Stafford
R. Stapleton
D. Starret
J. Stein
R. Stern
R. Stevenson
M. Stewart
J. Stiglitz
A. Stockman
D. Suits
L. Summers
M. Sumner
J. Sweeney
P. Taubman
J. Taylor
W. Thomson
R. Toikka
R. Tollison
R. Townsend
J. Triplett
J. Underwood
D. Usher
J. Vanderweide
B. Vann
R. Van Order
H. Varian
R. Verrecchia
D. Vining
P. Wachtel
M. Wachter
T. Wales
F. Warren-Boulton
R. Waud
L. Wegge
L. Weiss
L. W. Weiss
F. Welch
F. Westfield
J. Whalley
L. White
L. Wilde
J. Williams
R. Willis
R. Wilson
D. Wise
K. Wolpin
R. Wong
G. Wood
J. Wood
F. Wykoff
H. Yoshikawa
L. Young
E. Zajac
V. Zarnowitz
D. Zilberman

GEORGE H. BORTS, *Managing Editor*

Report of the Managing Editor

Journal of Economic Literature

As this is the last annual report I shall submit as the managing editor of the *Journal of Economic Literature*, I choose to take this occasion to bring to the attention of the membership some observations that I have formulated during the more than twelve years I will have served in this post.[1] The foundation issue was dated March 1969 (in fact, it was slow in the preparation and came from the printer closer to June). The last issue for which I will take responsibility will be the March 1981 issue. In all I will have organized and seen through the press forty-nine issues. Seven times seven seems a fortuitous number. In addition, we have produced eight annual *Indexes of Economic Articles*.

As is my tradition, I reproduce several tables intended to show not only the scope of the most recent year's issues (January through December 1980) but also to show in some general way what is to me a rather large block of work as well as a significant segment of my career as a professional economist.

Table 1 illustrates the projected allocation of space in the *Journal of Economic Literature* for 1980 as well as comparisons with the years 1978 and 1979 and totals for the period 1969 through 1980. Table 2 classifies the material by subject, both for the 1980 issues and for the total period. And, finally, Table 3 classifies the material by technical difficulty.

My staff has been particularly helpful. Several will continue to work for Moses Abramovitz, my successor, when there is a devolution. The associate editor (Naomi Perlman), the assistant editor (Drucilla Ekwurzel), the principal secretary (Lyndis Rankin), and the journals' secretary (Margaret Yanchosek) will continue to process the book annotations and the journals, including classification and abstracts, in Pittsburgh. This staff will also continue to pro-

[1] A large part of the observations in the original Report, as delivered, appears in *JEL*, March 1981, pp. 1–4; it is not reproduced here.

TABLE 1—QUANTITATIVE ANALYSIS OF *JEL* CONTENTS, 1978-80, AND TOTAL, 1969-80
(Number of pages in parentheses)

	1978		1979		1980		Total, 1969-80	
	No.	Pages	No.	Pages	No.	Pages	No.	Pages
Survey articles	4	(180)	4	(175)	4	(143)	41	(1455)
Essays on subfields	4	(79)	8	(147)	4	(103)	57	(1183)
Review articles	1	(12)	–	–	4	(28)	27	(219)
Articles about economic literature	2	(39)	1	(15)	6	(61)	16	(227)
Communications	1	(3)	–	–	7	(30)	65	(294)
Books annotated	1200	(259)	1201	(258)	1201	(272)	14,482	(2795)
Books reviewed	182	(286)	166	(293)	188	(294)	2,116	(3197)
Journal issues listed and indexed	921	(180)	962	(187)	1020	(201)	10,748	(1939)
Number of individual articles	7344	–	7437	–	8082	–	76,697	–
Subject index of journal articles	–	(360)	–	(377)	–	(438)	–	(3796)
Abstracts of articles	1649	(338)	1645	(336)	1714	(353)	18,448	(4162)
Total pages[a]		(1873)		(1877)		(2049)		(20,297)

[a] Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc.

TABLE 2—CLASSIFICATION BY SUBJECT, 1969-80

		1980		
			Creative	1969-80
		Commissioned	Curmudgeon	All articles
		Surveys	Essays	Total[a]
01	General	–	1	11
02	Theory	3	–	36
03	Thought (Methodology)	–	8	34
04	Economic History	–	–	4
05	Comparative Systems	–	–	4
11–12	Growth & Development	–	–	7
13	Stabilization	–	1	3
21–22	Econometric, Statistical Theory, Statistics	–	1	4
31	Monetary Economics	–	1	7
32	Fiscal Economics	–	–	7
40–44	International Economics	–	–	11
50	Managerial Economics	–	–	1
60	Industrial Organization, Industrial Regulation	–	2	3
70	Agricultural and Resource Economics	–	–	2
80	Labor Economics	1	–	9
90	Applied Welfare Economics Regional Economics	–	–	7
	TOTALS	4	14	150

[a] Includes all review articles on books, general essays on all literature

TABLE 3—CLASSIFICATION BY TECHNICAL DIFFICULTY, 1969-80

		1980	1969-80
		Creative	Totals Surveys; Creative
		Curmudgeon	Curmudgeon Articles;
	Surveys	Articles	Others[a]
Most Difficult	3	–	26
Some Difficulty	1	8	66
Not Difficult	–	6	58
TOTALS	4	14	150

[a] Review articles on books and general essays on all literature; excludes very short communications.

duce the annual *Index of Economic Articles*. Their offices will be associated with Carnegie-Mellon University. Madeline Fichter, this year my own secretary, will stay with me at the University of Pittsburgh. All five of these women have been splendid collaborators. The Association and I were lucky to have them.

I must thank the many persons who have served on the *Journal's* Board of Editors:

Moses Abramovitz
Irma Adelman
Marcus Alexis

Nicholas W. Balabkins
William Baumol
Abram Bergson

Martin Bronfenbrenner
Edwin Burmeister
Anne P. Carter

Richard A. Easterlin
Solomon Fabricant
David I. Fand
George R. Feiwel
Karl A. Fox
Alexander Gerschenkron
Arthur S. Goldberger
Marshall McGowan Hall
Arnold C. Harberger
Donald J. Harris
Hendrik S. Houthakker
D. Gale Johnson
John W. Kendrick
Peter B. Kenen
James K. Kindahl
Anne O. Krueger
David E. Laidler
Robert J. Lampman
Harvey Leibenstein
Thomas Mayer
Daniel L. McFadden
Allan H. Meltzer
William H. Miernyk
Michio Morishima
Marc L. Nerlove
Harry T. Oshima
Michael J. Piore
Roger L. Ransom
Barbara B. Reagan
Ryuzo Sato
Isabel V. Sawhill
Tibor Scitovsky
William Vickrey
E. Roy Weintraub
Charles Z. Wilson
Arnold Zellner

I wish also to thank the Chancellor of the University of Pittsburgh, Dr. Wesley Posvar, and the Dean of the Faculty of Arts and Sciences, Dr. Jerome Rosenberg, for the help they have given the Association, the *Journal*, and me. It has been a vast undertaking for the University of Pittsburgh, costly in space and services. And while there are many people at the University of Pittsburgh during the dozen years I have produced the *JEL* there who deserve thanks, I particularly wish to express my appreciation to the Provost-Emeritus, Dr. Charles Peake, and to a long-time secretary for the *Journal*, June Cox.

MARK PERLMAN, *Managing Editor*

Report of the Director

Job Openings for Economists

During 1980, employers advertised 2,051 new vacancies, a record number. Of these 1,416 (69 percent) were classified as academic and 635 (31 percent) were nonacademic. Last year employers advertised 1,928 new vacancies; 67 percent were academic and 33 percent nonacademic. This division between academic and nonacademic of roughly two-to-one has been the same for several years. Table 1 shows total listings (employers), total vacancies, new listings and new vacancies, by type for each issue of *JOE* in 1980.

TABLE 1—JOB LISTINGS FOR 1980

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	123	255	94	179
April	79	130	75	124
June	46	61	41	54
August	55	132	49	126
October	137	353	118	319
November	103	262	103	262
December	208	560	125	352
Subtotal	751	1,753	605	1,416
Nonacademic				
February	22	64	19	49
April	35	140	32	129
June	25	94	21	79
August	24	91	20	75
October	30	146	24	120
November	40	182	25	121
Subtotal	198	779	163	635
TOTALS	949	2,532	768	2,051

Universities with graduate programs and four-year colleges continue to be the major sources of job listings. They constitute 48 and 31 percent, respectively, of total employers. This compares to last year's 43 and 34 percent for the two. Table 2 shows the number of employers by type for each 1980 issue. The distribution is similar to that in 1976, 1977, 1978, and 1979.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economist that employers are seeking. The applied area of specialization seems to be of secondary importance. Table 3 shows the number of citations by field of specialization. General economic theory (000) led, followed by monetary and fiscal (300), econometrics and statistics (200), and welfare and urban (900). This pattern is also the same as that of the past several years.

The proposed 1981 budget and the 1980 (adopted and estimated) and 1979 (adopted and actual) budgets are given in Table 4. The 1980 approved budget projected a deficit (including indirect costs) of $15 thousand. The estimated actual deficit is $14 thousand. Total revenues are expected to be $21.6 thousand, total direct costs $13.6 thousand, and total indirect costs $22 thousand. The proposed budget for 1981 projects revenues of $21 thousand, total direct costs of $14 thousand, and total indirect costs of

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1980

Issue	Four-Year Colleges	Universities with Graduate Programs	Junior Colleges	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	44	79	–	7	2	1	1	7	4	145
April	41	38	–	9	4	5	3	9	5	114
June	20	26	–	5	3	2	1	9	5	71
August	19	36	–	3	3	4	–	10	4	79
October	46	91	–	7	3	4	–	14	2	167
November	38	65	–	6	4	3	1	8	–	125
December	83	125	–	14	5	4	–	15	2	248
TOTALS	291	460	–	51	24	23	6	72	22	949

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1980

Field[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	116	75	41	61	156	113	241	803
Growth and Development (100)	25	33	14	22	38	14	49	195
Econometrics and Statistics (200)	55	48	25	34	71	38	94	365
Monetary and Fiscal (300)	60	33	18	34	74	50	117	386
International Economics (400)	27	20	10	15	31	30	57	190
Business Administration, Finance, Marketing and Accounting (500)	37	29	11	19	49	30	69	244
Industrial Organization (600)	36	27	15	19	37	34	84	252
Agriculture and Natural Resources (700)	29	29	17	22	35	19	49	200
Labor (800)	27	17	14	15	41	22	64	200
Welfare and Urban (900)	38	23	7	21	48	38	78	253
Related Disciplines (A00)	10	5	4	4	10	5	12	50
Administrative Positions (B00)	9	2	8	8	17	6	16	66
TOTALS	469	341	184	274	607	399	930	3,204

[a] Fields of specialization codes are from the *Journal of Economic Literature.*

TABLE 4—JOB OPENINGS FOR ECONOMISTS BUDGET FOR 1981 (IN THOUSANDS)

	1978		1979		1980		1981
	(Adopted)	(Actual)	(Adopted)	(Actual)	(Adopted)	(Estimated)	(Proposed)
Revenue:							
Subscriptions		16.0		19.2		20.6	
Miscellaneous		.3		.2		1.0	
Total Revenue	$20	$16.3	$20	$19.4	$20.0	$21.6	$21.0
Expenses:							
Direct:							
Computer		1.5	2	.5	.5	.7	.8
Typewriter Rental		.8	2	1.5	2	1.4	1.5
Postage		3.9	4	4.8	5	4.7	5.0
Printing		4.4	4	5.7	5	6.5	6.0
Miscellaneous		.4	1	.1	.5	.3	.7
Total direct		10.9	14	12.6	13	13.6	14.0
Indirect:							
Salaries		17.5	16	21.0	22	22	25.0
Other							
Total Indirect		17.5	16	21.0	22	22	25.0
Total Expenses	$27	$28.4	$30	$33.6	$35	35.6	39.0
SURPLUS (DEFICIT)	(7)	(12.1)	(10)	(14.2)	(15)	(14)	(18.0)

$25 thousand. This leads to a projected accounting deficit of $15 thousand.

Violet Sikes continues to do virtually all the work involved in the publication and distribution of *JOE*. I wish to express my gratitude to her for a splendid job.

C. Elton Hinshaw, *Director*

The Committee on the Status of Women in the Economics Profession

Using data from the CSWEP roster, this report relates some of the current realities about the population of women economists and about their status. It finds that women are more professional, more research-oriented, and more diverse in their areas of specialization than legend would have it. It is not legend, however, that the status of women economists is and remains poor in academe, particularly in the prestige universities.

The pool of women economists in the CSWEP roster stood at 1,705 women in April 1980. This represents about 10 percent of total AEA membership. The report reveals that the vast majority are not merely occasional members of the work force—some 90 percent of nonstudents in the CSWEP roster work full time. Nearly 90 percent have advanced degrees, with half the roster members having Ph.D.s. More than a third are known to have published one or more books and articles, thus putting to rest the allegation that women do not write. Nearly three-quarters of the roster members have a primary field of specialization in economics different from the manpower, labor, and welfare fields normally associated with women's issues. Thus, it is a myth that most women economists concentrate on women's studies. What is true is that the set of universities referred to as "The Chairperson's Group" has hired a disproportionately large number of women from those fields, some 44 percent.

Half the women in the CSWEP roster have chosen academic careers. Roughly one-fifth have chosen government service or service in nonprofit sector, and another one-fifth work in the industrial, banking, or consulting sectors. For the academic sector, the roster provides evidence that the number of women economists may be nearly twice as large as had previously been supposed. Nevertheless, the proportion of women in high faculty positions has actually worsened relative to that of men, both at all institutions combined and at the major universities. For example, whereas the proportion of full professorships among male economists in tenure track positions at the Chairperson's Group increased from 51 to 57 percent from 1972 to 1978, that of women economists was only 26 percent in 1972 and fell even further to 18 percent in 1978.

There continues to be virtually no representation of women at the tenured level in the top seven economics departments. Thus, whereas the male economists who have achieved national reputations have done so with the powerful economics departments as their springboard, women economists have not been given this opportunity. Had it not been for government, which has in recent years done remarkably well, there would have been no improvement in the status of women economists since CSWEP was founded nearly a decade ago.

I. CSWEP Committee Activities

Before turning to the details of the analysis, I will summarize some of the activities of CSWEP since the December 1979 meeting at Atlanta. At that meeting, the principle of nondiscrimination on the basis of sex was affirmed by the decision of the Executive Committee to hold meetings and job markets through 1985 in states that have ratified the Equal Rights Amendment, and by a resolution of the general membership which applauded and affirmed the wisdom of that decision. A number of women aided in the formulation of the resolution, including representatives from URPE as well as past and present members of CSWEP. CSWEP had some $261 in member's donations left over from the ERA ad campaign: $100 of this was given to ERA Georgia in December, and the other $161 was used to support "A National ERA Evening," cochaired by Rosalynn Carter and Betty Ford, held in Washington in June. The proceeds from the evening are to be used by ERA America for citizen education and lobbying and by the National Women's Political Caucus ERA fund to support the election campaigns of key state legislators. Since the American Economic Association cannot supply funding directly or indirectly for such purposes, CSWEP was pleased when Heather Ross and Belle Sawhill supplied the additional funds to achieve the $500 required to have CSWEP named as a sponsor of the event.

A second major activity of CSWEP is the pending reorganization of the Committee, to adapt it to its growing role as an umbrella organization for women involved in the regional economics associations. CSWEP is designating four committee members to represent the four regions: East, West, Midwest, and South. Each regional member will in turn appoint a three- or four-person executive committee to support and lead CSWEP activities in the specified regions. Each regional CSWEP group will coordinate with the president and officers of the regional economics associations to plan CSWEP activities at their meetings. These activities will include a session on research related to women's issues, a CSWEP business meeting, and a social get together. A representative of the recently formed Washington Women Economists (WWE) will also sit in on CSWEP committee meetings. WWE was formed early in 1979 as a network for women economists living and working in the nation's capitol. WWE arranges many programs, including conferences, dinner meetings, etc., publishes a bimonthly newsletter and a membership directory, services job inquiries, and encourages research and student activities. Since CSWEP has not in the past paid a great deal of attention to the needs of the women economists working in government, it is wonderful indeed that the WWE group has begun to fill this important role.

The initial leaders of the regional activities will be Irma Adelman (agricultural economics, Berkeley), Chair, CSWEP-West assisted by Claire Vickrey (economics, Berkeley), Myra Strober (education, Stanford), and Sara Bechman (California State Government). Heading CSWEP-South and Southwest is Joan Haworth (economics, Florida State) assisted by Ruth Andress (business, University of South Carolina), Mary Fish (business, University of Alabama), Persis Rockwood (marketing, Florida State) and Judy Pitcher (Consumer Product Safety Commission). Janet Goulet (business, Wittenberg), is the Chair, CSWEP-Midwest, assisted by Kim Sosin (economics, University of Nebraska), and George Thoma (economics, Elmherst). Jean Shackelford (economics, Bucknell) heads CSWEP-East, assisted by Teresa Amott (Wellesley), Judith Stitch (American Council on Education), and Julianne Malveaux (management, New School).

The Denver AEA meetings featured Alice M. Rivlin, Director of the Congressional Budget Office of the U.S. Congress, as the speaker at a joint CSWEP-American Finance Association luncheon. Her talk described the profound changes in the government budgeting process that have taken place over the past five years, and gave a preview of improvements that are now under consideration. There were also two major sessions of particular interest to women. One was the traditional CSWEP research session with the topic being the effect of inflation on labor force participation and the distribution of household income. A second CSWEP-sponsored session described some proven techniques for improving the status of women in all types of employment: academic, business, government, and labor. This session was particularly lively. Mentoring and networking techniques were advocated, as well as selection by women of fields such as micro-economic theory and econometrics where demand is strong. In addition, the importance of commitment from the top was stressed as a key ingredient to improving status. Thus, direct techniques must be supplemented by efforts to affect the decisions of persons in top positions.

Finally, Marianne Ferber has carried out an analysis of the use of the CSWEP roster. One of the resolutions adopted when CSWEP was formed required the provision of a roster of its women members, listing their qualifications and fields, which was to be made widely known to all prospective employers. Consequently, CSWEP began to compile a roster which, by the end of 1973, contained 1,400 names. This number has grown somewhat to 1,705 names. From the beginning, the roster has been used as a mailing list for the CSWEP *Newsletter*, and this continues to be one of its functions.

The primary purpose of the roster, provision of a list of women economists along with information about them that is useful to potential employers, committees seeking qualified women to serve on panels, boards, etc., has also been served since CSWEP has made the roster available for a nominal charge, which helps defray the cost of the operation.

Unfortunately, the roster has had only limited use. The number of requests received was 17 in 1976, 18 in 1977, 10 in 1978, and 14 in 1979. In an attempt to improve usage, CSWEP has revised the format of the roster to make it more legible, has made it more up-to-date by switching from an annual to a semi-annual up-date, and has sent a mailing to academic institutions and some government agencies telling them of the availability of the roster. The result of all these changes has been a more than 100 percent increase in requests. But the total number is still only 31 for January–May 15, 1980.

CSWEP is interested in exploring new and easier ways of getting information to potential users. We are currently exploring ways to improve the computer accessibility of the data. We have also begun to use the data for analysis. Any suggestions for further improvements in the production or use of the roster or for additional ways to bring its existence to the attention of potential users will be gratefully received.

II. Analysis of the Status of Women Economists

Because of the timing of the 1980 meeting of the American Economic Association, the Universal Academic Questionnaire data used for reporting purposes by my predecessors are not available. Thus, my report must rely on other data sources. Fortunately, the CSWEP roster has recently been put into analyzable format by the able and energetic efforts of Marianne Ferber. A computer program to analyze these data is being prepared by Joan Haworth, who plans to use it in her own research and to make it available for future CSWEP reports. Since this program is not yet complete, Beverly Loudermilk and Anna Pegram of my office undertook the tedious task of collating at least some of the roster data, which they did with remarkable care and good cheer. Using letters and phone calls, they also assisted in compiling a list of women economists who are assistant, associate, and full professors at the Chairperson's Group of Universities.

Tables 1–3 summarize in tabular form information contained in the April 1980 CSWEP roster. Table 1 displays the distribution of women economists by highest degree. It is seen that nearly 90 percent of the women on the CSWEP roster have advanced degrees. The distribution of advanced degrees across primary field of specialization indicates that the proportion of women with only a Bachelors' degree is significantly higher than average in only two fields, economic statistics and business and finance, where about 20 percent of women economists have only Bachelors' degrees. Slightly more than a quarter of women economists have a Masters as their highest degree. The economic statistics and business and finance fields are again the two fields in which this proportion is significantly above the average. Thus, women in these two business-related fields tend to stop their education sooner than do women in the more academic fields of specialization in economics.

Roughly half the women in the roster have completed their Ph.D.s. This contrasts with about 13 percent who have all but their doctorates. Thus, nearly four times more women economists have continued their education through the Ph.D. level than have quit before completing their dissertations. Since a number of roster members are still students, even this may overvalue the rate of noncompletion. I cannot but believe that a comparison of these figures with those of the AEA membership as a whole would reveal that the record of women is no worse than that of men in this regard. Certainly, our figures dispel the notion that most women economists tend to drop out rather

TABLE 1—PERCENTAGES OF WOMEN ECONOMISTS BY HIGHEST DEGREE AND EMPLOYMENT STATUS

	Highest Degree					Status			
Primary Field of Specialty in Economics	B.A., B.S.	M.A., M.S.	A.B.D.	Ph.D.	Other	Employed Full Time	Student	Employed Part Time	Other
000 General Economics	7.6	29.8	9.7	52.5	0.4	84.9	6.7	5.5	2.9
100 Economic Growth, Development	11.0	28.9	8.7	51.4	0.0	75.7	11.0	4.6	8.8
200 Economic Statistics	20.9	34.2	9.0	35.1	0.8	74.6	13.4	6.0	6.0
300 Monetary and Fiscal	8.4	24.3	17.8	48.6	1.0	78.5	14.5	5.1	1.9
400 International Economics	10.9	25.6	16.0	46.2	1.3	78.2	13.5	5.1	3.1
500 Business and Finance	20.1	40.2	5.8	33.3	0.6	90.2	3.5	2.9	3.5
600 Industrial Organization	12.7	27.1	18.6	40.7	0.8	80.5	12.7	4.2	2.5
700 Agriculture	10.6	23.4	12.8	51.1	2.1	85.1	6.4	6.4	2.1
800 Manpower, Labor	6.3	18.0	17.7	56.9	1.2	82.0	10.6	3.9	3.5
900 Welfare Programs	7.7	23.0	15.3	53.1	1.0	80.1	7.7	5.1	7.1
TOTAL	10.9	27.2	13.3	47.9	0.8	81.0	10.0	4.8	4.2

Source: CSWEP Roster, April 1980.

than to complete their dissertations. Preliminary analysis of the roster data also reveals that significantly higher percentages of women are attaining the Ph.D. as their highest degree in the years since 1970 than in prior years.

Table 1 shows that over 80 percent of the women in the roster are employed full time. Ten percent are students, 5 percent are employed part time, and 4 percent have other or unknown status. These data suggest that of the women who care enough about maintaining a professional link to have joined CSWEP, only a small minority have selected part-time employment or have dropped out of the labor market. The vast majority are serious members of the labor force.

Table 2 displays the type of employment chosen by women in the CSWEP roster. Roughly half have chosen academic careers. In all but two fields (general economics, and business and finance), one-fourth to one-fifth have chosen government service or service in the nonprofit sector. While the

TABLE 2—PERCENTAGES OF WOMEN ECONOMISTS BY TYPE OF EMPLOYMENT AND JOB AVAILABILITY

	Employment				Availability for Other Positions			
Primary Field of Specialty in Economics	Academic	Government/ Nonprofit	Indus./ Banking Consulting	Other/ Unknown	Actively Looking	Consider Good Offer	Not Interested New Position	Unknown
000 General Economics	77.3	3.4	8.8	10.5	5.9	41.2	22.3	30.7
100 Economic Growth, Development	37.6	19.1	26.6	16.7	10.4	54.3	11.0	24.3
200 Economic Statistics	35.8	20.1	27.6	16.4	9.0	45.5	20.9	24.6
300 Monetary and Fiscal	52.8	22.4	10.7	14.0	11.2	47.2	19.2	22.4
400 International Economics	37.2	24.4	21.8	16.7	16.0	44.2	18.6	21.4
500 Business and Finance	33.3	9.2	48.9	8.6	11.5	44.8	16.7	27.0
600 Industrial Organization	38.1	23.7	27.1	11.0	7.6	60.2	10.2	22.0
700 Agriculture	55.3	21.3	17.0	6.4	6.4	55.3	12.8	25.5
800 Manpower, Labor	56.1	21.6	12.9	10.2	8.6	53.7	18.0	19.6
900 Welfare Programs	51.5	20.9	14.3	13.3	11.2	45.4	23.5	19.9
TOTAL	49.3	17.8	20.4	13.8	9.9	48.3	18.1	23.6

Source: CSWEP Roster, April 1980.

TABLE 3—PERCENTAGES OF WOMEN ECONOMISTS BY PUBLICATION RECORD AND BY ACADEMIC RANK

Primary Field of Specialty in Economics	Articles and Books			Women Academics by Academic Rank				
	None	One or More	Unknown	Dean/ Dept. Head	Prof.	Assoc. Prof.	Asst. Prof.	Instructor, Etc.
000 General Economics	35.3	35.3	29.4	5.3	18.0	23.3	27.5	25.9
100 Economic Growth, Development	35.8	38.2	26.0	0.0	20.0	20.0	30.8	29.2
200 Economic Statistics	55.2	20.9	23.9	4.3	10.9	17.4	21.7	45.6
300 Monetary and Fiscal	42.5	32.7	24.8	3.5	15.9	21.2	29.2	30.1
400 International Economics	41.7	31.4	26.9	7.0	15.8	21.1	21.1	30.3
500 Business and Finance	51.2	23.6	25.3	5.2	24.1	19.0	19.0	32.7
600 Industrial Organization	50.8	28.0	21.2	4.4	15.6	11.1	28.9	39.9
700 Agriculture	31.9	29.8	38.3	0.0	11.5	23.1	23.1	42.4
800 Manpower, Labor	27.8	41.6	30.6	2.1	18.9	23.1	30.8	25.2
900 Welfare Programs	27.6	46.9	25.5	3.0	22.8	14.9	31.7	27.6
TOTAL	39.0	34.2	26.8	3.7	18.1	20.3	27.6	30.3

Source: CSWEP Roster, April 1980.

overall average of women with jobs in the business sector is about one-fifth of the roster population, the distribution ranges from a low of less than 10 percent in the general economics area to a high of nearly 50 percent in the business and finance area. Slightly less than 14 percent of roster members did not fill in employment information. Primarily these were women who had indicated student or other status, and hence the employment question was not pertinent for them. Certainly, Table 2 suggests that more effort should be expended by CSWEP on behalf of its members who have chosen nonacademic careers.

Table 2 also reveals that women economists are a great deal more flexible about considering job changes than is generally supposed. Only 18 percent of the women in the roster have indicated they are not interested in a new position. Almost 50 percent would consider a good job offer, and another 10 percent are actively looking for a job.

Table 3 reveals that more than one-third of the roster members have published one or more books and articles in contrast to just under 40 percent who are known to have no publications. While the known publishers thus do not form the majority of our membership, the proportion of publishing women is certainly high enough to dispel the notion that women tend to teach but not to publish. Not surprisingly, the two fields where women had the least proportion of advanced degrees, economic statistics and business and finance, also display the lowest publication rates. The two fields with the highest proportion of publishing women are the fields of manpower, labor, and welfare. Thus, women who are interested in fields where women's issues play a major role tend to write somewhat more than women with other primary fields of specialization.

Table 3 reveals that approximately 70 percent of women academics are in tenure track positions. About 38 percent have achieved appointments at the associate professor level or above. To see whether this is an improvement in status, I compared this distribution with that reported in the first annual CSWEP report of May 1973. At that time approximately 79 percent of the women were in tenure track positions, and 38 percent had achieved appointments at the associate professor level or above. So women's status has not improved in this regard. Table 4 displays a number of other comparisons over time and over data sets of women's status. The only improvement is that the roster data indicates that women have achieved somewhat higher ranks on average than have been reported by the Universal Academic Questionnaire data. In other respects, both for universities as a whole, and for the

TABLE 4—FACULTY DISTRIBUTION OF TENURE TRACK RANKS BY SEX

	Universal Academic Questionnaire, 1972				Universal Academic Questionnaire, 1978-79				CSWEP Roster, April 1980	
	Females		Males		Females		Males		Females	
	No.	Distribution	No.	Distribution	No.	Distribution	No.	Distribution	No.	Distribution
	All Colleges and Universities, Economics Departments									
Full Professors	48	21.5	1489	38.2	48	19.7	1637	43.4	184	31.3
Associate Professors	59	26.5	1055	27.1	58	23.8	1005	26.6	171	29.1
Assistant Professors	116	52.0	1350	34.7	138	56.5	1130	30.0	233	39.6
TOTAL	223		3894		244		3772		588	
	Chairperson's Group of Universities,[a] Economics Departments								CSWEP Survey March 1980	
Full Professors	14	26.4	563	51.4	10	17.5	653	57.4	20	22.1
Associate Professors	8	15.1	211	19.3	8	14.0	184	16.2	17	16.8
Assistant Professors	31	58.5	321	29.3	39	68.4	301	26.4	58	61.1
TOTAL	53		1095		57		1138		95	

[a]The Chairperson's Group of Universities was comprised of 43 major universities in 1972, and of 65 major universities in 1978–79.

Chairperson's Group as will be discussed later, women's status has declined in comparison with their earlier status and in comparison with that of men.

The most startling difference between the roster figures and those reported by the Universal Academic Questionnaire concern total numbers of women in academe. The 1972–73 questionnaire data cited 223 women economists in tenure track positions. Five years later those data showed a slight increase to 244 women. The CSWEP roster reveals some 843 academic women; 588 of the women are in tenure track positions, over twice as many as have been picked up using the traditional questionnaire as the data source. Thus, the roster data reveal that there have been serious underestimates in the pool of women economists. The discrepancy is due in part to the fact that many universities do not complete the questionnaire data. In addition, because the questionnaire data are directed only at economics departments, substantial numbers of women economists who are in other departments are not picked up.

The status of women economists in the Chairperson's Group of Universities has always been considered an important indicator of our stature in the profession. With this in mind and because of doubts about the completeness of the questionnaire data, CSWEP conducted a survey this spring of the women economists at the Chairperson's Group of Universities. A total of 95 women in economics departments were found who were at a rank of assistant professor or above. This figure is nearly double the 57 women reported in the 1978–79 Universal Academic Questionnaire data. The survey data are reasonably complete for the economics departments of the major campuses, and include at least some women economists from other departments and campuses. Indeed, the survey identified an additional 44 women economists in the Chairperson's universities whose appointments were in business schools, in departments of city or regional planning, in agricultural economics departments, in outlying campuses, and so forth. This represents a 40 percent increase in the number of women economists who would otherwise have shown up in the Chairperson's group.

Aside from finding a larger pool of women in these universities, the survey is not very heartening. Within the 65 economics departments of the major campuses of the Chairperson's group, 40 had no tenured women professors, 18 had one tenured woman, and 7 had two tenured women. Moreover, of the 40 departments with no tenured women, 17 also had no untenured women. So even using the survey data, there is no doubt that these departments remain nearly totally male.

The distribution according to professional rank of women in the Chairperson's economics departments is also disheartening. As Table 4 shows, whereas only 40 percent of tenure track women in the CSWEP roster are assistant professors, 61 percent of the women in the Chairperson's group of economics departments are at this rank. The associate professors constitute 17 percent of the Chairperson's departments, but 29 percent in the membership at large. Full professors are 31 percent of the roster population, but only 22 percent of the Chairperson's group. A comparison of women's distribution by rank with that of the men in these departments is equally disheartening. In the 1973 CSWEP report, 51 percent of the male professors in the Chairperson's economic departments were full professors as contrasted to 26 percent of the women. In the 1980 CSWEP report, 57 percent of the men were full professors as contrasted to only 18 percent of the women. According to the CSWEP survey, 22 percent of the women professors in these universities are full professors. So in this dimension, as well, our status appears to have slipped rather than improved over the years since CSWEP was founded.

In the top seven economics departments, only one, MIT, has a tenured woman. There are no tenured women in the economics departments of Harvard, Yale, Princeton, Chicago, Stanford, or Berkeley. Both MIT and Berkeley do, however, have a tenured woman in a noneconomics department. Thus, the traditionally dismal record of the top economic departments in the nation remains dismal as far as women are concerned.

Although the prestige schools have been closed to women, government service has opened up during 1973–80. Not surprisingly, many of the women in our profession have risen to prominance along this latter path. Unlike academe, women economists have been appointed to the highest government offices: to the Cabinet; to the Council of Economic Advisors; to Commissionerships as varied as the Federal Reserve Board, the Bureau of Labor Statistics, the Civil Aeronautics Board, and the Consumer Product Safety Commission; to directorships as varied as those at the Congressional Budget Office and the National Commission for Employment Policy. I conducted in 1980 a count of women economists who were at supergrade and appointed positions in Washington and found at least as many women there as are tenured in economics departments at the Chairpersons Group of Universities. Thus, although there are substantially fewer numbers of women economists in government, they have achieved relatively greater stature than their counterparts in academe.

Table 5 displays primary field of specialization among four groups of women economists—the entire CSWEP roster, the members of the CSWEP roster who are still students, the women economists in all departments at the Chairperson's Group of Universities, and the women who have joined the Washington Women Economists group. The first column shows the percentages for the AEA membership as a whole. The CSWEP distribution does not look that different from the distribution of the total AEA membership. The only categories that differ by more than three percentage points are manpower, and labor and welfare where the proportion of women are higher. It is interesting that the women students in these categories are at roughly the same level as the total AEA membership. The new crop of women economists differ from their predecessors in focusing their attention away from women's issues and toward areas such as industrial organization and international economics.

Some 44 percent of the women who have found jobs in the Chairperson's Group of Universities are, surprisingly, in the two women-related issue codes of manpower, and labor and welfare. The manpower, labor figure for the women in the Chairpersons' Group is nearly double that for CSWEP women as a whole. Thus, the impression in the Chairperson's Group of Universities that most women economists work in fields related to women's issues seems to arise because these universities have recruited

TABLE 5—PERCENTAGES OF VARIOUS GROUPS OF ECONOMISTS BY PRIMARY FIELD OF SPECIALIZATION

		Women Economists			
Primary Field of Specialty in Economics	Total AEA Membership	CSWEP Roster	CSWEP Students	Chairperson's Group	WWE
000 General Economics	17	14	7	16	2
100 Economic Growth, Development	12	10	11	6	12
200 Economic Statistics	8	8	13	5	9
300 Monetary and Fiscal	14	13	14	11	9
400 International Economics	9	9	14	8	12
500 Business and Finance	9	10	4	2	4
600 Industrial Organization	9	7	13	8	7
700 Agricultural	6	3	6	1	10
800 Manpower, Labor	9	15	11	28	15
900 Welfare Programs	8	12	8	16	20

Source: 1978 AEA *Directory of Economists*, CSWEP Roster, April 1980, CSWEP Survey of Chairperson's group, March 1980, Washington Women Economists Membership Directory, 1980.

such women more than they have women from other fields.

The fields of specialization of the 273 members of Washington Women Economists are distinctive in that there are few generalists and larger percentages of women in the fields of development, international, agriculture and welfare than in the Chairpersons' group. Other interesting statistics (not displayed on the chart) are that the WWE group has a lower proportion of Ph.D.s (32 percent vs. 48 percent) than the CSWEP roster as a whole, and that the distribution of WWE membership by type of employment yields 53 percent working in government, 33 percent in business, 10 percent in academe, and 4 percent not employed.

ELIZABETH E. BAILEY, *Chair*

Report of the Representative to the National Bureau of Economic Research

During 1980 work continued on the several large-scale research projects initiated by the National Bureau of Economic Research in 1979. These include the project on the changing role of debt and equity finance in the United States directed by Benjamin Friedman, the inflation project directed by Robert Hall, the analysis of private and public pensions directed by John Shoven with the assistance of Zvi Bodie and Kim Clark, the simulation of the effects of changes in tax policy under the direction of Martin Feldstein, and the cooperative effort with economists in Germany, Sweden, and the United Kingdom in the production of comparable figures on capital taxation in the four countries. The NBER survey of black youths was completed and Richard Freeman's Labor Studies group began analysis of the data in an effort to determine the causes and consequences of large-scale unemployment among this group.

Richard Freeman also launched a new Bureau program of research on productivity, compensation and employment which is designed to study some of the major institutional and structural changes in *U.S.* labor market institutions such as trade unions, governmental regulations, and corporate personnel policy, as well as trends such as the increasing average age of the work force.

Bureau conferences (and organizers) in the United States and abroad in 1980 included "Postwar Changes in the American Economy" (Martin Feldstein); "Import Competition and Adjustment" (Jagdish Bhagwati); "Economic Aspects of Health" (Victor Fuchs); "Inflation" (Robert Hall); "Trade Prospects Among the Americas" (in cooperation with the Fundacao Instituto de Pesquisas Economicas, Sao Paulo, and the Illinois Bureau of Business and Economic Research), (Malcom Gillis for the NBER); "International Seminar in Macroeconomics" (Robert Gordon and Georges de Menil); and "Economics of Compensation" (Sherwin Rosen). It is expected that volumes or special issues of journals will result from the first five conferences listed, and most of the papers from these conferences are or will be available in the NBER Conference Papers series.

In 1980 the following bureau books were published by the University of Chicago Press: *Population and Economic Change in Developing Countries*, Richard Easterlin, ed.; *New Developments in Productivity Measurement*, John W. Kendrick and Beatrice N. Vaccara, eds.; *The Measurement of Capital*, Dan Usher, ed.; *Rational Expectations and Economic Policy*, Stanley Fischer, ed.; *Modeling the Distribution and Intergenerational Transmission of Wealth*, James D. Smith, ed.; *The American Economy in Transition*, Martin Feldstein, ed.; and *Doctors and Their Workshops: Physician Influences on the Use of Medical Resources*, Mark Pauly. From the Ballinger Publishing Company came: *Commodity Markets and Latin American Development—A Modeling Approach*, Walter C. Labys, M. Ishaq Nadiri, and Jose Nunez del Arco, eds.; and *Business Cycles, Inflation and Forecasting*, Geoffrey H. Moore.

The Bureau's Business Cycle Dating Group identified January 1980 as the most recent peak in *U.S.* business activity. The group further stated that unless there were an extraordinarily sharp and quick reversal of activity, this peak would mark the onset of a recession.

Effective July 1, Michael Boskin succeeded Robert Michael as Director of the Bureau's Palo Alto Office.

During July and August 1980 a Summer Institute was held at the Bureau's Cambridge office. Workshops were held in Financial Markets and Monetary Economics, Taxation, Pensions, Labor Economics, Productivity and Technical Change, and International Studies. About 150 researchers took part in the Institute for differing lengths of time.

Further information on Bureau activities is available in the NBER *Reporter*, from

Charles E. McLure, Jr., Vice President, NBER, 1050 Massachusetts Avenue, Cambridge, Massachusetts 02138, or from the undersigned at Johns Hopkins University.

CARL F. CHRIST, *Representative*

Report of the Committee on U.S.–Soviet Exchanges

Following the fifth U.S.–Soviet Economic Symposium held in the United States in June 1979, the Soviet Economic Association invited ten American economists to a sixth symposium in the U.S.S.R. in 1980. The October dates suggested, however, were not feasible in terms of American university teaching schedules, so the symposium was postponed until the following year. It is now scheduled to be held in Alma-Ata, U.S.S.R., June 1981. The conference theme is "The Role of the State in Price Formation." It is expected that the *U.S.* and Soviet participants will present parallel papers in such areas as agricultural pricing, energy pricing, and overall price stabilization. After the conference discussions, the American delegates are invited to visit universities and research institute in Tashkent, Moscow, and Leningrad.

LLOYD G. REYNOLDS, *Chair*

Report of the Committee on Economic Education

The revision of the Test of Understanding College Economics (*TUCE*), a standardized instrument designed to evaluate knowledge at the principles level, was completed during 1980. This project, under the direction of Phillip Saunders (Indiana University), Rendigs Fels (Vanderbilt), and Art Welsh (Joint Council on Economic Education), has involved (1) the development of an improved test-question classification scheme, (2) the construction of several principles tests (macro, micro, macro-micro combined), (3) the gathering of test norming data with the cooperation of approximately thirty colleges and universities, and (4) the analysis of the norming data. A paper describing the results of the test norming was presented at the Denver meetings and is published elsewhere in this issue. The revised *TUCE* is now in publication. It will be distributed by the Joint Council on Economic Education, 1212 Avenue of the Americas, New York, New York 10036. Saunders is presently undertaking research using the norming data to provide additional data on the efficacy of the test under alternative conditions.

The second of the Lilly-sponsored workshops for disseminating the Teacher Training Program (*TTP*) was held at the University of Wisconsin-Madison under the directorship of W. Lee Hansen. The *TTP* is a packaged course, complete with *Resource Manual* and video tapes, designed to train graduate students in the techniques of teaching economics. The summer workshops, including the one held at Madison, train faculty in the use of the *Resource Manual* and in the structuring of *TTP*'s to be organized at their home institutions. The Wisconsin program drew fifty participants from thirty-two colleges and universities. In addition to Hansen, workshop leadership was provided by Michael Salemi (University of North Carolina-Chapel Hill), Phillip Saunders (Indiana University), and Art Welsh (Joint Council on Economic Education). Guest lectures were provided by Darrell Lewis (Minnesota), John Siegfried (Vanderbilt) and Jeffrey Wolcowitz (Harvard). Due in part to the demonstrated success of the Wisconsin workshop, as well as a similar forum held at Indiana University the previous summer, the Lilly Endowment has recently announced the funding of two additional workshops for the summers of 1981 and 1982. Information on the *Resource Manual* and the *TTP* can be obtained from the Joint Council on Economic Education.

The project to study the status of the economics major, under the leadership of John Siegfried (Vanderbilt) and funded by the Sloan Foundation, is in the midst of its data collection phase. This involves taking a census of economics departments to identify the nature of their majors and course offerings. Data will additionally be gathered directly from graduating majors at twenty-five to fifty colleges and universities. Siegfried will develop a summary report for the Washington AEA meetings which describes the attributes of majors, the nature of course offerings and requirements, the reasons for concentration, the forms of instructional technology, and the postgraduate plans of majors. This AEA presentation will be followed by a panel discussion on the major. It is expected that, based on the Siegfried report, the Committee on Economic Education will undertake future programming on the major. Most of its programming to date has concentrated on the principles course.

ALLEN C. KELLEY, *Chair*

Report of the Economic's Institute Policy and Advisory Board

The Economic's Institute had another successful year in 1980 with a total of 532 participants, or a 24 percent increase in separate enrollments over 1979. Enrollments by sessions increased as follows: 29 percent for the spring session, 17 percent for the summer session, and 40 percent for the fall session.

With the year-around program well established, increasing numbers of Institute students are taking longer periods of preparatory training, since the Institute can now accommodate lower beginning levels of proficiency in English, and also provide more adequate overall preparation both in English and in core subject areas. Of the 532 participants in 1980, the attendance of only 183 or 34 percent was limited to the Institute's traditional five or ten weeks summer session programs. The remainder completed more extended preparatory programs with 23 percent terminating their programs at the Institute in August, and 27 percent terminating in December for January university admissions.

The Institute is now also involved in placing a larger number of its students at universities subsequent to their enrollment at the Institute. Most of these students, whose initial selection has been based on other criteria than pre-existing proficiency in English, are proving to be exceedingly promising, and, in this way, the Institute is helping to broaden the foreign student selection process without sacrificing quality. It expects to become a more important source of high-quality students in the years ahead and is expanding cooperative relations with an increasing number of universities that are admitting students directly from the Institute with excellent results.

Work has begun on the sixth edition of the Institute's *Guide to Graduate Study in Economics and Agricultural Economics in the United States and Canada.* This biennial edition will be published by the Richard D. Irwin Company near the end of 1981.

The Institute was separately incorporated as a nonprofit educational institution during the year, and the Policy and Advisory Board now serves as the Institute's Board of Directors. Carlos Diaz Alejandro, Yale University, and Raymond Vernon, Harvard University, completed three-year terms on the Board during the year. Bent Hansen, University of California-Berkeley, and Louis Wells, Harvard University, are new three-year appointees. The Board held a regular meeting in Boulder on July 26, 1980, with a follow-up meeting in September at the Allied Social Science Association annual meeting in Denver.

EDWIN S. MILLS, *Chair*

The Choices of Power

Utilities Face the Environmental Challenge

Marc J. Roberts and Jeremy S. Bluhm

The Tennessee Valley Authority. Ontario Hydro. The Pacific Gas and Electric Company. The Southern Company. Boston Edison. The Los Angeles Department of Water and Power. Richly detailed case studies of these six electric utilities reveal strikingly different ways of dealing with regulations to protect the environment, and offer a framework for understanding the decision-making process in large firms.
8 flowcharts 7 maps 12 halftones $25.00

The Economics of Justice

Richard A. Posner

Richard Posner applies economic theory to the theory of justice, primitive and ancient social and legal institutions, the law and economics of privacy and reputation, and the law and economics of racial discrimination. A major theme is that the logic of the law in many ways appears to be an economic one, and judges, for example, in interpreting the common law, act as if they were trying to maximize economic welfare. $25.00

Taking Your Medicine

Drug Regulation in the United States

Peter Temin

Peter Temin's timely book describes how the government has taken control over the sale and use of medicinal drugs.

"Choices about medicines used to be made by consumers. Temin shows how the choice has shifted in turn to doctors, then to experts, and now to bureaucrats who assume that people cannot and should not make choices for themselves. Must reading for anyone who disagrees with a Big Daddy form of government."
—*Louis Lasagna, M.D.,* School of Medicine and Dentistry, University of Rochester
$18.50

A Theory of Price Control

John Kenneth Galbraith

With a new introduction by the author

Mr. Galbraith discusses why price controls were used successfully during World War II. Many of the lessons learned from this wartime experience could be applied to our current economic difficulties.

"This witty and trenchant book is as relevant today as it was almost 30 years ago. It is the best thing of its kind."—*Robert L. Heilbroner*
$2.95, paper $7.95, cloth

Hospital Costs and Health Insurance

Martin S. Feldstein

Why has the cost of hospital care grown so rapidly in the past thirty years? In this pathbreaking book, Martin Feldstein sets forth a novel explanation and proposes a new national health insurance plan—one that would give incentives to physicians and their patients to economize on costs and simultaneously provide adequate coverage for major illnesses.
$25.00

The Telecommunications Industry

The Dynamics of Market Structure

Gerald W. Brock

"A stunning piece of work... no other broad based study of the telecommunications industry can be remotely compared with this one."
—*Richard E. Caves*

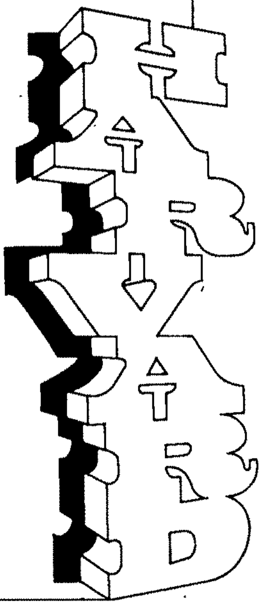
The telecommunications industry—normally a fast changing one—is currently undergoing even faster changes, spurred by rapid technological progress and a changing regulatory environment. Brock's study of the evolving structure of the industry focuses on A.T.&T.
Harvard Economic Studies, 151 $25.00

Equality, the Third World, and Economic Delusion

P.T. Bauer

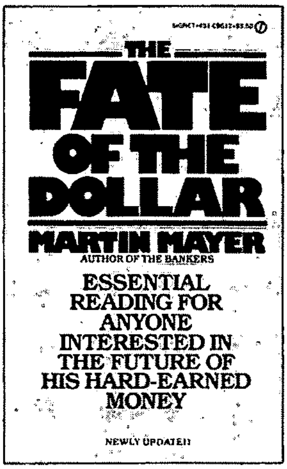
"There are few books in economics which are a pleasure to read. Bauer's is one of the few."—*A.A. Walters*

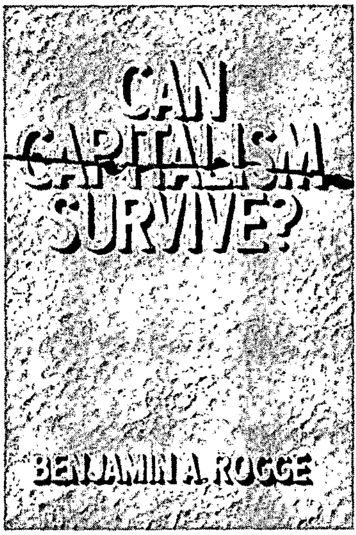
Even in impoverished countries lacking material and human resources, Bauer argues, economic growth is possible under the right conditions. The author challenges many views about development that are widely held in this important and controversial book. $17.50

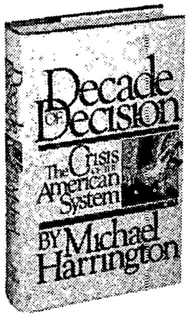


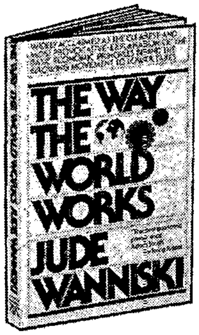
Harvard University Press

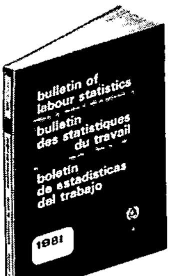
Cambridge, Massachusetts 02138

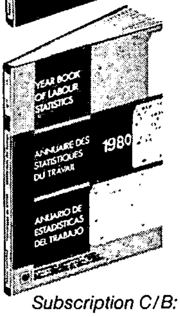


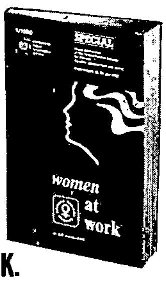

CAN CAPITALISM SURVIVE?
BENJAMIN A. ROGGE











An indispensable tool for:
Economists Teachers Students
Researchers Authors
Reference Librarians

Index of Economic Articles

prepared under the auspices of

The Journal of Economic Literature

of the

American Economic Association

Each volume in the *Index* lists, both by subject and by author, articles in major economic journals and in collective volumes published during a specific year. The *Index* volumes also include articles or testimony from selected congressional hearings or government documents published during the year. Many of the features in the *Index* cannot be found in any other single reference source.

Currently available are:

Volume	Year Covered
XI	1969
XII	1970
XIII	1971
XIV	1972
XV	1973
XVI	1974
XVII	1975
XVIII	1976

Future volumes will be published regularly to keep the series as current as possible.

Price: $50.00 per volume (special 30% discount to A.E.A. members)

Distributed by

RICHARD D. IRWIN, INC.
Homewood, Illinois 60430

Note: Volumes I through X, VI-A, and VII-A, previously published by Irwin, are available at $7.95 per copy.

Begin making plans to attend the

Ninety-Fourth

Annual Meeting of The American Economic Association

(in Conjunction with Allied Social Science Associations)

to be held in

WASHINGTON, D.C.

Dec. 28-30, 1981

The Employment Center opens Sunday, December 27.

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

The 1982 meeting will be held in New York, NY, December 28-30.

TECHNICAL CHANGE AND ECONOMIC POLICY

AMERICAN ECONOMIC ASSOCIATION

Organized at Saratoga, New York, September 9, 1885

PAST OFFICERS

Presidents

Francis A. Walker,* M.I.T., 1886–92
Charles F. Dunbar,* Harvard, 1893
John B. Clark,* Columbia, 1894–95
Henry C. Adams,* Michigan, 1896–97
Arthur T. Hadley,* Yale, 1898–99
Richard T. Ely,* Wisconsin, 1900–01
Edwin R. A. Seligman,* Columbia, 1901–03
Frank W. Taussig,* Harvard, 1904–05
Jeremiah W. Jenks,* Cornell, 1906–07
Simon N. Patten,* Pennsylvania, 1908
Davis R. Dewey,* M.I.T., 1909
Edmund J. James,* Illinois, 1910
Henry W. Farnum,* Yale, 1911
Frank A. Fetter,* Princeton, 1912
David Kinley,* Illinois, 1913
John H. Gray,* Minnesota, 1914
Walter F. Willcox,* Cornell, 1915
Thomas N. Carver,* Harvard, 1916
John R. Commons,* Wisconsin, 1917
Irving Fisher,* Yale, 1918
Henry B. Gardner,* Brown, 1919
Herbert J. Davenport,* Cornell, 1920
Jacob H. Hollander,* Johns Hopkins, 1921
Henry R. Seager,* Columbia, 1922
Carl C. Plehn,* California, 1923
Wesley C. Mitchell,* Columbia, 1924
Allyn A. Young,* Harvard, 1925
Edwin W. Kremmerer,* Princeton, 1926
Thomas S. Adams,* Yale, 1927
Fred M. Taylor,* Michigan, 1928
Edwin F. Gay,* Harvard, 1929
Matthew B. Hammond,* Ohio State, 1930
Ernest L. Bogart,* Illinois, 1931
George E. Barnett,* John Hopkins, 1932
William Z. Ripley,* Harvard, 1933
Harry A. Millis,* Chicago, 1934
John M. Clark,* Columbia, 1935
Alvin S. Johnson,* New School, 1936
Oliver M. W. Sprague,* Harvard, 1937
Alvin H. Hansen,* Harvard, 1938
Jacob Viner,* Chicago, 1939
Frederick C. Mills,* Columbia, 1940
Sumner H. Slichter,* Harvard, 1941
Edwin G. Nourse,* Brookings, 1942
Albert B. Wolfe,* Ohio State, 1943
Joseph S. Davis,* Stanford, 1944
I. L. Sharfman,* Michigan, 1945
E. A. Goldenweiser,* Inst. Advanced Study, 1946
Paul H. Douglas,* Chicago, 1947
Joseph A. Schumpeter,* Harvard, 1948
Howard S. Ellis, California, 1949
Frank H. Knight,* Chicago, 1950
John H. Williams, Harvard, 1951
Harold A. Innis,* Toronto, 1952
Calvin B. Hoover,* Duke, 1953
Simon Kuznets, Pennsylvania, 1954
John D. Black,* Harvard, 1955
Edwin E. Witte,* Wisconsin, 1956
Morris A. Copeland, Cornell, 1957
George W. Stocking,* Vanderbilt, 1958
Arthur F. Burns, Columbia, 1959
Theodore W. Schultz, Chicago, 1960
Paul A. Samuelson, M.I.T., 1961
Edward S. Mason, Harvard, 1962
Gottfried Haberler, Harvard, 1963
George J. Stigler, Chicago, 1964
Joseph J. Spengler, Duke, 1965
Fritz Machlup, Princeton, 1966
Milton Friedman, Chicago, 1967
Kenneth E. Boulding, Colorado, 1968
William J. Fellner, Yale, 1969
Wassily Leontief, Harvard, 1970
James Tobin, Yale, 1971
John Kenneth Galbraith, Harvard, 1972
Kenneth J. Arrow, Harvard, 1973
Walter W. Heller, Minnesota, 1974
Robert Aaron Gordon,* California, 1975
Franco Modigiliani, M.I.T., 1976
Lawrence R. Klein, Pennsylvania, 1977
Jacob Marschak,* California, 1978†
Tjalling C. Koopmans, Yale, 1978
Robert M. Solow, M.I.T., 1979
Moses Abramovitz, Stanford, 1980

Secretaries

Richard T. Ely,* 1886–92
Edward A. Ross,* 1893
Jeremiah W. Jenks,* 1894–96
Walter F. Willcox,* 1897–99

Treasurers

Edwin R. A. Seligman,* 1886–90
Frederick B. Hawley,* 1891–95
Charles H. Hull,* 1896–99

Secretary-Treasurers

Charles H. Hull,* 1900
Frank A. Fetter,* 1901–06
Winthrop M. Daniels,* 1907–08
Thomas N. Carver,* 1909–13
Allyn A. Young,* 1914–20
Ray B. Westerfield,* 1921–25
Frederick S. Deibler,* 1925–35
James Washington Bell,* 1936–61
Harold F. Williamson, 1962–70
Rendigs Fels, 1971–75

Editors

Davis R. Dewey,* 1911–40
Paul T. Homan,* 1941–51
Bernard F. Haley, 1952–62
John G. Gurley, 1962–68
Arthur Smithies, 1963–68
George H. Borts, 1969–80
Mark Perlman, 1969–80

*Deceased
†Died before taking office

The American Economic Review

ARTICLES

JUNE 1981

THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members six times a year, in March, May, June, September, and semimonthly in December.

Dues for 1981, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$30.00 for regular members with rank of assistant professor or lower, or with annual income of $14,400, or less;

$36.00 for regular members with rank of associate professor, or with annual income of $14,400 to $24,000;

$42.00 for regular members with rank of full professor, or with annual income above $24,000;

$15.00 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are $100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $5.00 to cover extra postage.

●Correspondance relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

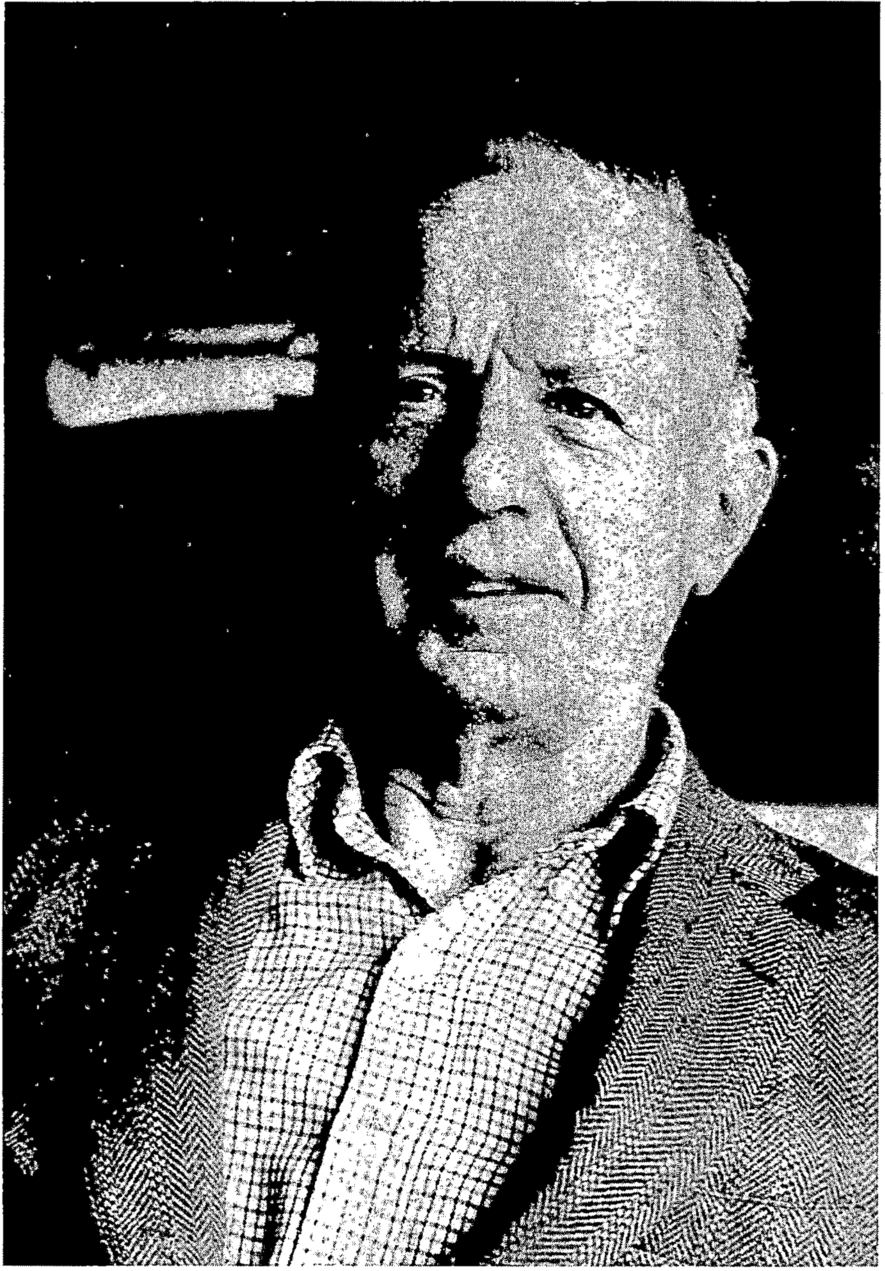
●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212

CHARLES P. KINDLEBERGER

DISTINGUISHED FELLOW
1981

Charles P. Kindleberger's contributions to international economics and to economics education have taken many forms. Initially, for more than a decade, he performed as an able public servant. In his later role as extraordinary teacher, he has trained a large proportion of the next generations' leading figures in his field. His sympathetic and active involvement with the education of minority students led to a year's teaching at Clark College and a long association as trustee of the Associated Negro Colleges of Atlanta.

As a scholar and governmental adviser he has been prodigiously creative and productive: at the forefront of the intense postwar discussions concerned with the form and implications of the international monetary system, with balance-of-payment issues, with policies for trade and economic growth, with foreign-investment problems, and with international migration. His work is always marked by imaginative and penetrating insights, informed by a profound historical knowledge of the way institutions and people behave, and fortified by impressive judgment. His views have been cogent, influential, and fruitful; his policy prescriptions have been relevant and challenging.



Charles P. Kindleberger

THE AMERICAN ECONOMIC REVIEW

●Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

●No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

June 1981

VOLUME 71, NUMBER 3

Articles

Shorter Papers

A Modigliani-Miller Theorem for Open-Market Operations

By NEIL WALLACE*

Monetary policy determines the composition of the government's portfolio. Fiscal policy, in particular, the size of the deficit on current account, determines the path of net government indebtedness. In this paper I will show that alternative paths of the government's portfolio consistent with a single path of fiscal policy *can be* irrelevant in precisely the sense in which the Modigliani-Miller theorem shows that alternative corporate liability structures are irrelevant. Irrelevance here means that both the equilibrium consumption allocation and the path of the price level are independent of the path of the government's portfolio. Roughly speaking, the irrelevance proposition I prove has the following form: if there is an equilibrium with certain properties for one path of portfolios for the government, then that equilibrium is also an equilibrium for a large class of other paths of portfolios for the government provided only that lump sum taxes are adjusted in an appropriate way. Appropriate means, among other things, that fiscal policy is held constant.

I prove the irrelevance result for a limited class of environments: models of two-period-lived, overlapping generations with a single consumption good that is storable via a constant returns to scale, stochastic technology. This class of models, described in Section I, is broad enough to include examples that establish the nonvacuousness of the "if" clause of the irrelevance proposition. Nonvacuousness requires that there be equilibria in which the private sector *voluntarily* holds real capital and unbacked government liabilities, liabilities that I call fiat money. The overlapping-generations structure and risky real capital (storable consumption good) make this possible.

In Section II, I describe the conditions for a perfect-foresight, competitive equilibrium for the Section I environments. The irrelevance proposition is presented and proved in Section III. The proposition establishes conditions under which the amount of the consumption good purchased by the government in the open market for fiat money and stored by the government is irrelevant. Complete markets in contingent claims play a prominent role just as they do in Joseph Stiglitz' version of the Modigliani-Miller theorem. In Section IV, to help interpret the irrelevance proposition, I discuss the relationship between the fiscal policy assumptions used to obtain irrelevance and assumptions that are usually used to characterize an unchanged fiscal policy.

In Sections V and VI, I consider by way of examples departures from the assumptions of the irrelevance proposition. In the former, I describe a departure that arises when the nonnegativity restriction on private gross investment is binding. In the latter, I describe a departure that arises when a legal restriction on minimum money holdings is binding. These examples establish the necessity of the voluntarily diversified asset-holding assumptions of the irrelevance proposition. They are also of interest because they offer a micro-economic interpretation of the usual macro-economic model analysis of open-market operations.

I. The Physical Environment

Time is discrete and there is a single good. At each date t, a new generation of $N(t)$ two-period-lived individuals (generation t)

*University of Minnesota and the Federal Reserve Bank of Minneapolis. Financial support was provided by NSF grant, SOC 77-22743, and by the Federal Reserve Bank of Minneapolis. The views expressed herein are solely my own and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. I am indebted to colleagues for helpful comments on an earlier draft.

appears. Each member h of generation t maximizes the expected value of $u^h(\ , \ )$, where the first (second) argument is consumption of the good by h in the first (second) period of life and where u^h is strictly increasing, strictly concave, and twice differentiable.

At each date t, there is a new aggregate endowment of $Y(t)>0$ units of the consumption good. This good may be consumed or stored. If $K(t)\geqslant 0$ is the aggregate amount placed into storage at t, then $K(t)x/(t+1)+xY(t+1)$ is the total amount available at $t+1$, where $x(t+1)$ is a random variable drawn independently from period to period from a discrete probability distribution: $x(t+1)=x_i>0$ with probability f_i; $i=1,2,\ldots,I$. The I-element vector $(x_1,x_2,\ldots,x_I)$ will be denoted x. The value of $x(t+1)$ is observed after time t storage is determined and before generation $t+1$ appears. Note that $K(t)$ is the sum of nonnegative private storage, $K^p(t)$, and nonnegative government storage, $K^g(t)$.

The supply of fiat money is determined by the government. Changes in it do not require the expenditure of resources by the government and private storage of fiat money neither affects its physical properties nor requires the expenditure of resources.

II. The Market Scheme

I will describe the conditions for a perfect foresight, competitive equilibrium in terms of time t markets for claims on time $t+1$ consumption in "state" $x(t+1)=x_i$. The members of generation t in their role as consumers demand such claims. Firms, owned by members of generation t in their role as producers, supply such claims by storing the consumption good and by storing fiat money. In general, the government announces a policy, including a lump sum tax-transfer scheme, in terms of such claims.

A. *The Consumer's Lifetime Choice Problem*

The consumer choice problem of the young of generation t is described in terms of the following notation:

$(c_1^h(t), c_2^h(t))=$ the $(I+1)$-element consumption vector of member h of generation t where $c_1^h(t)$ is first-period consumption, $c_2^h(t)=(c_{21}^h(t), c_{22}^h(t),\ldots,c_{2I}^h(t))$, and $c_{2i}^h(t)$ is second-period consumption in state $x(t+1)=x_i$.

$(w_1^h(t), w_2^h(t))=$ the corresponding $(I+1)$-element endowment vector of member h of generation t, where $w_2^h(t)=(w_{21}^h(t), w_{22}^h(t),\ldots,w_{2I}^h(t))$.

$s(t)=$ the I-element vector $(s_1(t), s_2(t),\ldots,s_I(t))$ where $s_i(t)$ is the price at time t of one unit of $t+1$ consumption in state $x(t+1)=x_i$ in units of time t consumption.

Later it will be convenient to have a notation for the consumption allocation and endowment of generation t. Thus, let $c(t)$ be the $N(t)(I+1)$-element vector consisting of one $(c_1^h(t), c_2^h(t))$ vector for each member h of generation t and let $w(t)$ be the corresponding $N(t)(I+1)$-element endowment vector.

This notation is meant to allow for possible dependence of, say, $s(t)$ on $x(t)$, $x(t-1)$, and so on. For any variable $\cdot(t)$, dependence on t is used to denote possible dependence on $x(t-j)$, $j\geqslant 0$. This is a convenient notation because the young of generation t make choices having observed $x(t-j)$, $j\geqslant 0$.

Member h of generation t is assumed to choose a nonnegative vector $(c_1^h(t), c_2^h(t))$ to maximize $\Sigma_i f_i u^h[c_1^h(t), c_{2i}^h(t)]$ subject to

$$(1)\quad c_1^h(t)+s(t)c_2^h(t)\leqslant w_1^h(t)+s(t)w_2^h(t)$$

where the vector multiplication is inner-product multiplication. For $s(t)$ and $(w_1^h(t), w_2^h(t))$ that imply a nonempty, bounded budget set, there is a unique maximizing vector $(c_1^h(t), c_2^h(t))$ given by the unique solution to equation (1) at equality and

$$(2)\quad f_i u_2^h\left[c_1^h(t), c_{2i}^h(t)\right] = s_i(t)\sum_{j=1}^{I} f_j u_1^h\left[c_1^h(t), c_{2j}^h(t)\right] \qquad i=1,2,\ldots,I$$

This is all that need be said about consumer demand.

B. *The Choice Problem of Firms*

In their role as producers, members of generation t may enter one or both of two lines of business at time t: storing the consumption good or storing money. In each line, any producer maximizes profit as a price taker with regard to $s(t)$ and the time t, and time $t+1$ prices of money.

Profit in terms of time t consumption from storing $k \geqslant 0$ units of the consumption good is $s(t)xk-k$. Since this is linear in k, the condition that storage be finite in any equilibrium implies as an equilibrium condition

$$s(t)x \leqslant 1 \tag{3}$$

a condition that must hold with equality if total private storage $K^p(t)$ is positive.

If $p(t)$ is the price of a unit of money at time t in units of time t consumption and $p(t+1)$ is the price of a unit of money at time $t+1$ in terms of time $t+1$ consumption (an I-element vector as of time t), then profit in terms of time t consumption from storing $m \geqslant 0$ units of fiat money is $s(t)p(t+1)m-p(t)m$. Since this is linear in m, finiteness of the supply of money implies that prices in any competitive equilibrium satisfy

$$s(t)p(t+1)=p(t) \tag{4}$$

We may write equality here because if firms store no money, then demand falls short of supply and $p(t)=0$.

C. *Government Policy Rules*

Government policy is a specification at time $t=1$ after $x(1)$ has been observed of paths, possibly contingent, for government consumption at t, $G(t) \geqslant 0$; the endowment vector for generation t, $w(t)$; government storage at t, $K^g(t)$; and the money supply at t, $M(t) \geqslant 0$. For $t \geqslant 1$ and each $x(t)$ in x, these are chosen subject to

$$K^g(t)+G(t)=T(t)+K^g(t-1)x(t) + p(t)\left[M(t)-M(t-1)\right] \tag{5}$$

Here T(t), total lump sum taxes minus transfers at t, is defined by [1]

$$T(t)=Y(t)-\sum_h w_1^h(t) - \sum_h w_{2i}^h(t-1) \tag{6}$$

and $M(0)$, $w(0)$ (the endowment of the old at $t=1$), and $K^g(0)$ are assumed given as initial conditions. (The summations over h are over the members of generation t and $t-1$, respectively, a convention that will be used throughout.)

D. *Perfect Foresight Competitive Equilibrium*

The question of foresight arises with regard to $p(t+1)$ in (4) and with regard to $w_2^h(t)$ in (1). Perfect foresight requires that the ith element of $p(t+1)$ in (4) be equal to the equilibrium price of money at $t+1$ in state $x(t+1)=x_i$ and that the $w_2^h(t)$ vector on the basis of which h chooses at t be realized at $t+1$. Put formally, then, for specified government policy consisting of a possibly contingent sequence $(G(t), w(t), K^g(t))$ defined for $t \geqslant 1$, a perfect foresight competitive equilibrium consists of nonnegative sequences $c(t-1)$, $s(t)$, $K(t) \geqslant K^g(t)$, $p(t)$ and $M(t)$ that for all $t \geqslant 1$ satisfy (1) at equality and (2) for each h, (3)–(6), and

$$\sum_h \left(c_{2i}^h(t)-w_{2i}^h(t)\right) = K^p(t)x_i + p_i(t+1)M(t) \tag{7}$$

for each i. The left-hand side of (7) is the aggregate excess demand of consumers for consumption at $t+1$ in state $x(t+1)=x_i$, while the right-hand side is the supply of such consumption by firms in that state.

E. *An Example*

Here is an example of an economy with a diversified equilibrium:

Physical environment: For all t, $N(t)=N$, $Y(t)=yN>0$ and $u^h(z_1, z_2)=ln(z_1)+ln(z_2)$ for all h; $x=(x_1, x_2)=(0.5, 2.0)$ and $f_1=f_2=0.5$.

Policy: For all $t \geqslant 1$, $G(t)=K^g(t)=0$ and $w_1^h(t)=y$, $w_{2i}^h(t)=0$ for all i and h.

[1] That (5) with $T(t)$ as defined in (6) is the cash-flow constraint for the government, may, perhaps, be made more obvious by noting that it follows from feasibility (at equality) and a consolidated cash-flow constraint for the private sector. Letting $C(t)$ be total private consumption at t, the first of these is $G(t)+C(t)+K(t) = x(t)K(t-1)+Y(t)$ while the second is $C(t)+K^p(t) + p(t)M(t) = x(t)K^p(t-1) + p(t)M(t-1) + \Sigma_h w_1^h(t) + \Sigma_h w_{2i}^h(t-1)$. Equation (5) is obtained by subtracting one of these from the other.

Equilibrium: For all $t \geqslant 1$, $(s_1(t), s_2(t)) = (2/3, 1/3)$, $K(t)/N = y/4$, $M(t) = M(1)$, $p(t)M(t)/N = y/4$ and for all h, $(c_1^h(t), c_{21}^h(t), c_{22}^h(t)) = (y/2, 3y/8, 3y/4)$.

Thus, this economy has an equilibrium in which the money supply and the price of money are unchanging from $t=1$ on. Each young person consumes $y/2$ when young. Per capita saving, $y/2$, is composed of contingent claims on second-period consumption supported by a per capita portfolio consisting of real money balances equal to $y/4$ and storage of the consumption good equal to $y/4$. I will refer to this example and to closely related examples below.

III. The Irrelevance Proposition

The proposition to be proved is as follows:

PROPOSITION 1: *If* $\{\bar{c}(t-1), \bar{s}(t), \overline{K}(t), \bar{p}(t), \overline{M}(t)\}$ *is an equilibrium with* $\bar{p}(t) > 0$ for all $t \geqslant 1$ *for the policy* $\{G(t), w(t), K^g(t)\} = \{\overline{G}(t), \overline{w}(t), 0\}$, *then* $\{\bar{c}(t-1), \bar{s}(t), \overline{K}(t), \bar{p}(t), \hat{M}(t)\}$ *is an equilibrium for the policy* $\{\overline{G}(t), \hat{w}(t), \hat{K}^g(t)\}$, *where* $\{\hat{K}^g(t)\}$ *is any nonnegative sequence bounded by* $\{\overline{K}(t)\}$ *and* $\{\hat{w}(t)\}$ *is any* $w(t)$ *sequence that for all* $t \geqslant 1$ *satisfies*

(a) $\hat{w}_1^h(t) + \bar{s}(t)\hat{w}_2^h(t) = \overline{w}_1^h(t) + \bar{s}(t)\overline{w}_2^h(t)$ *for each* h, *and*

(b) $\Sigma_h[\hat{w}_{2i}^h(t) - \overline{w}_{2i}^h(t)] = \hat{K}^g(t)[x_i - \bar{p}_i(t+1)/\bar{p}(t)]$ *for each* i. (The notation "$\{\cdot(t)\}$" means a sequence defined for all $t \geqslant 1$.)

Before giving a proof, it is worth noting that the proposition is not vacuous. Nonvacuousness requires that there exist economies having equilibria with $p(t) > 0$ for all t and $\overline{K}(t) > 0$ for at least some t when $K^g = 0$. The example given at the end of section two meets this requirement.

Nonvacuousness also requires the existence of at least one $\{\hat{w}(t)\}$ that satisfies (a) and (b). One such sequence is given by

$$\tilde{w}_1^h(t) = \overline{w}_1^h(t);$$

$$\tilde{w}_{2i}^h - \overline{w}_{2i}^h(t)$$
$$= \hat{K}^g(t)\big[x_i - \bar{p}_i(t+1)/\bar{p}(t)\big]/N(t)$$

for all h, i, and $t \geqslant 1$. The $\{\tilde{w}(t)\}$ endowment scheme obviously satisfies (b). To show that it satisfies (a), note that

$$\bar{s}_i(t)\big[\tilde{w}_{2i}^h(t) - \overline{w}_{2i}^h(t)\big]$$
$$= \hat{K}^g(t)\big[\bar{s}_i(t)x_i - \bar{s}_i(t)\bar{p}_i(t+1)/\bar{p}(t)\big]/N(t)$$

Summing both sides over i, we obtain

$$\bar{s}(t)\big[\tilde{w}_2^h(t) - \overline{w}_2^h(t)\big]$$
$$= \hat{K}^g(t)\big[\bar{s}(t)x - 1\big]/N(t) = 0$$

where the first equality follows from (4) and the second from the fact that $\tilde{K}^g(t) > 0$ implies $\overline{K}(t) > 0$, and hence, (3) at equality. This and $\tilde{w}_1^h(t) = \overline{w}_1^h(t)$ imply that $\{\tilde{w}(t)\}$ satisfies condition (a).

PROOF:

By condition (a), if $\bar{c}(t)$, $\bar{s}(t)$, and $\overline{w}(t)$ satisfy (1) at equality and (2), then so do $\bar{c}(t)$, $\bar{s}(t)$, and $\hat{w}(t)$. Moreover, since (3) and (4) hold at equality at the prices $\bar{s}(t)$ and $\bar{p}(t)$, all that remains is to show that (7) is satisfied by the $\hat{M}(t)$ implied by (5) with $p(t) = \bar{p}(t)$. First note that

$$\Sigma_h\big[\hat{w}_1^h(t) - \overline{w}_1^h(t)\big] = 0 \qquad \text{for all } t \geqslant 1 \tag{8}$$

To derive this, multiply (b) by $\bar{s}_i(t)$ and sum over i. Then use (3) and (4) at equality to obtain $\bar{s}(t)\{\Sigma_h[\hat{w}_2^h(t) - \overline{w}_2^h(t)]\} = 0$. This and (a) summed over h imply (8).

Now, in order to find $\hat{M}(t)$, subtract (5) for the $K^g(t) = 0$ policy from (5) for the $K^g(t) = \hat{K}^g(t)$ policy to get

$$\begin{aligned} \hat{K}^g(t) = {} & \hat{T}(t) - T(t) \\ & + \big[\hat{K}^g(t-1) - \overline{K}^g(t-1)\big]x(t) \\ & + \bar{p}(t)\big[\hat{M}(t) - \hat{M}(t-1) \\ & - \overline{M}(t) + \overline{M}(t-1)\big] \end{aligned} \tag{9}$$

Since $M(0)$, $K^g(0)$, and $w_{2i}(0)$ are fixed by initial conditions, (8) implies $\hat{T}(1) = \overline{T}(1)$.

Thus, for $t=1$, (9) reduces to

$$(10) \quad \hat{K}^g(t)=\bar{p}(t)\big[\hat{M}(t)-\overline{M}(t)\big]$$

I now show by induction that (10) holds for all $t\geqslant 1$. If (10) holds for some $t\geqslant 1$, then (9) for $t=\bar{t}+1$ is

$$(11) \quad \hat{K}^g(\bar{t}+1)=\hat{T}(\bar{t}+1)-\overline{T}(\bar{t}+1)$$
$$+\hat{K}^g(\bar{t})x(\bar{t}+1)$$
$$+\bar{p}(\bar{t}+1)\big[\hat{M}(\bar{t}+1)-\overline{M}(\bar{t}+1)\big]$$
$$-\hat{K}^g(\bar{t})\bar{p}(\bar{t}+1)/\bar{p}(\bar{t})$$

But, for all t,

$$\hat{T}(t+1)-\overline{T}(t+1)=\textstyle\sum_h\big[\bar{w}^h_{2i}(t)-\hat{w}^h_{2i}(t)\big]$$
$$=-\hat{K}^g(t)\big[x(t+1)+\bar{p}(t+1)/\bar{p}(t)\big]$$

where the first equality follows from (6) and (8) and the second from condition (b). Upon substituting this into (11), we get (10) for $t=\bar{t}+1$ as required.

I now use (10) to show that $\bar{c}(t+1)$, $\hat{w}(t)$, and $\hat{M}(t)$ satisfy (7). I begin with (7) for the "$-$" equilibrium, namely,

$$(12) \quad \textstyle\sum\bar{c}^h_{2i}(t)-\sum_h\bar{w}^h_{2i}(t)$$
$$=\overline{K}(t)x_i+\bar{p}_i(t+1)\overline{M}(t)$$

Upon substituting for $\sum_h\bar{w}^h_{2i}(t)$ from condition (b), we have

$$\textstyle\sum_h\bar{c}^h_{2i}(t)-\sum_h\hat{w}^h_{2i}(t)=\big[\overline{K}(t)-\hat{K}^g(t)\big]x_i$$
$$+\bar{p}_i(t+1)\big[\overline{M}(t)+\hat{K}^g(t)/\bar{p}(t)\big]$$

Finally, using (10), we get

$$(13) \quad \textstyle\sum_h\bar{c}^h_{2i}(t)-\sum_h\hat{w}^h_{2i}(t)$$
$$=\big[\overline{K}(t)-\hat{K}^g(t)\big]x_i+\bar{p}_i(t+1)\hat{M}(t)$$

This is (7) for the asserted equilibrium under the $\hat{K}^g(t)$ policy and completes the proof.

IV. An Unchanged Fiscal Policy

What is the relationship between the assumptions about fiscal policy used to prove irrelevance and assumptions that characterize an unchanged fiscal policy? Unchanged fiscal policy means (i) an unchanged path of government consumption, (ii) an unchanged distribution of income, and (iii) an unchanged path of total taxes minus transfers. (Given (i), (iii) is equivalent to holding the path of the deficit unchanged.) Since (i) is an irrelevance assumption and (ii) has its counterpart in condition (a), it remains to explore the relationship between (iii) and condition (b).

Since different paths of the government's portfolio imply different paths of net interest received by the government and since net interest received by government is a component of taxes minus transfers, (iii) requires that these net interest differences be offset by differences in other components of taxes minus transfers. In order to express this requirement, I need a definition of time t net interest or earnings on a government portfolio $(K^g(t-1), M(t-1))$ held from $t-1$ to t. I define time t earnings on this portfolio to be $[x(t)-1]K^g(t-1)-[p(t)-p(t-1)]M(t-1)$, a definition which includes the capital loss on government liabilities in the form of money. In terms of this definition, requirement (iii) for the two portfolios $(\overline{K}^g(t+1), \overline{M}(t-1))$ and $(\hat{K}^g(t-1), \hat{M}(t-1))$ at the prices $\bar{p}(t)$ is

$$(14) \quad \overline{T}(t)-\hat{T}(t)=\big[x(t)-1\big]$$
$$\times\big[\hat{K}^g(t-1)-\overline{K}^g(t-1)\big]$$
$$-\big[\bar{p}(t)-\bar{p}(t-1)\big]\big[\hat{M}(t-1)-\overline{M}(t-1)\big]$$

where the right-hand side is the difference in earnings (net interest) implied by the two portfolios.

As a preliminary step, it is convenient to show that (14) and (10) are equivalent. To show that (14) implies (10), note that (10) for $t=1$ follows from (9) for $t=1$, given initial conditions which by (14) imply $\bar{T}(1)=\hat{T}(1)$. The following induction argument establishes (10) for all t. Assume (10) holds for some $\bar{t}\geqslant 1$. Start with (9) for $t=\bar{t}+1$ and substitute into it (14) for $t=\bar{t}+1$ and (10) for $t=\bar{t}$. The result is (10) for $t=\bar{t}+1$. To show that (10) implies (14), simply substitute (10) for t and $t-1$ into (9). This works for $t\geqslant 2$. For $t=1$, use initial conditions in place of (10) lagged. This equivalence is not surprising since (10) expresses the requirement that the path of net government indebtedness at the prices $\bar{p}(t)$ be the same under the alternative portfolios, which is another way of saying that the path of the deficit be the same.

We can now establish some conclusions. Since (10) was shown to be a consequence of the assumptions used to prove irrelevance, it follows that those assumptions imply an unchanged path of fiscal policy in the sense of (i)–(iii). I have not, however, been able to establish whether the converse holds, or, in particular, whether (14) could replace condition (b) among the assumptions used to prove irrelevance. It is trivial to show that condition (b) is necessary in the sense that (14) and irrelevance imply (b).[2] But this necessity may be vacuous because I have not been able to determine whether (b) is implied by (14) and the other assumptions of the irrelevance proposition. It can, however, be shown that no "simple" endowment scheme satisfies (14) and condition (a), but not (b) at prices satisfying (3) and (4) at equality.

Finally, it is worth noting that condition (b) is reminiscent of a condition that holds automatically in expositions of the Modigliani-Miller theorem for corporate liability structures. Condition (b) requires that differences in earnings at t implied by different government portfolios held from $t-1$ to t be paid out in the form of taxes to agents who were present at $t-1$. For alternative corporate liability structures, this requirement is met automatically because poststate payouts by the corporation necessarily go to individuals who bought prestate titles to those payouts. Although holding fiscal policy unchanged in the sense of (i)–(iii) guarantees that differences in earnings are paid out through the tax system, given the overlapping-generations structure, (i)–(iii) alone may not guarantee that poststate payoffs go to individuals who were present prestate (see (14) and (6)).

V. Binding Nonnegativity of Private Storage

The bound on $K^g(t)$ is necessary in order to get irrelevance. If $K^g(t)>\bar{K}(t)$ for some t, then no feasible value of $K^p(t)$ consistent with unchanged total accumulation at t exists. Irrelevance cannot, then, hold because total resources at $t+1$ in each state depend on $K^g(t)$.

While this suffices to establish necessity of the bound on $K^g(t)$, I want to display an example which suggests that some features of the usually asserted effects of open-market operations are consistent with assuming that such operations occur in a range that violates the $K^g(t)\leqslant\bar{K}(t)$ bound. In particular, I will display an example in which $K^g(t)>\bar{K}(t)$ amounts to a subsidy on storage financed by lump sum taxes with the subsidy being greater and the price of money lower the greater is K^g.

The example is that given at the end of Section II except that I now assume $K^g(t)/Y(t)=\theta$, $w_{2i}^h(t)=\theta y[x(t+1)-1]$ for all h and $t\geqslant 1$ and $\Sigma_h w_{2i}^h(0)=K^g(0)x(1)$. For each θ in [0, 1/2], there is a stationary equilibrium with $p(t)=p_\theta>0$ for all $t\geqslant 1$.[3] For $\theta\leqslant 1/4$, the irrelevance proposition holds. For $\theta>1/4$, the stationary solution is found by first solving the relevant versions of (1), (2), (4), and (7) with $K^p(t)=0$ for c_1^h, c_{21}^h, c_{22}^h, s_1, s_2, and $(p_\theta M_\theta)$.[4] Then p_θ may be found using the relevant version of (5); namely,

$$(15)\quad \theta y=p_\theta M_\theta/N-(\bar{p}\bar{M}/N)(p_\theta/\bar{p})$$

$$=p_\theta M_\theta/N-(y/4)(p_\theta/\bar{p})$$

[2] Irrelevance implies that (12) and (13) hold. Subtract (13) from (12) and substitute into the difference $(\hat{M}(t)-\bar{M}(t))$ form (10). The result is condition (b).

[3] It is not true that 1/2 is an upper bound on θ. An (unattainable) upper bound is 2/3.

[4] The solution for s_1 is $[5\theta/2-2+2(1+\theta/2+73\theta^2)^{1/2}]/6\theta$. One must, of course, verify that the (s_1, s_2) solution satisfies (3). It does and with strict inequality when $\theta>1/4$. This, in turn, implies net taxes when $\theta>1/4$.

where $\bar{p}$ and $\overline{M}$ are the equilibrium values for $K^g=0$. Without displaying the numerical solutions, it can be shown that $p_\theta/\bar{p}<1$ for some θ.

In a stationary equilibrium for this economy

$$(16)\quad y=c_1(t)+K^g/N+(\bar{p}\overline{M}/N)(p_\theta/\bar{p})$$
$$=c_1(t)+\theta y+(y/4)(p_\theta/\bar{p})$$

This describes the disposition of the per capita endowment of the young at $t=1$, $p_\theta \overline{M}/N$ being the amount that goes to the current old. For this example, $c_1(t)$ is equal to half of wealth, which by (1) and (4) implies

$$(17)\quad c_1(t)=[y-\theta y(1-sx)]/2$$
$$\geqslant[y-\theta y/2]/2$$

The inequality follows from noting that sx is a minimum at $s_1=1$ for s satisfying (4). This inequality and (16) imply $p_\theta/\bar{p}\leqslant(2-3/\theta)$ or $p_\theta/\bar{p}<1$ for $\theta>1/3$.

An alternative way to generate stationary equilibria with $\theta>1/4$ is to treat $(p_\theta/\bar{p})$ as a policy instrument; the interpretation is that the government announces a price of money p_θ satisfying $0<p_\theta/\bar{p}<1$ at which it is willing to sell (or buy) money in exchange for the consumption good at any time. The equilibrium is found by solving (15) and the relevant versions of (1), (2), (4), (7), and $K^p(t)=0$ for θ, c_1^h, c_{21}^h, c_{22}^h, s_1, s_2, and $p_\theta M_\theta$.

There is, of course, nothing "neutral" about alternative values of $p_\theta/\bar{p}$ accomplished in either of these equivalent ways. The example shares features of an open-market operation consisting, say, of government purchases of mortgages on new construction in some locality at a higher-than-market price. In conformity with the example, this stimulates construction in the locality, and hence is not neutral. As is widely recognized, though, to call this monetary policy is stretching matters. First, fiscal policy cannot be held fixed in the face of this policy. Second, an equivalent policy is an interest subsidy which everyone agrees is fiscal policy.

VI. Globally Binding, Legal Minimum Money Holdings

The model described in Section II is one of voluntarily held money, equation (4) being a consequence. In fact, equation (4) is a consequence if *some* money is held voluntarily. But it is easy to construct a model in which money is held *only* to meet prescribed legal restrictions. In such situations, money can have value in an equilibrium with the left-hand side of (4) less than the right-hand side and irrelevance need not hold.

I will illustrate the nonirrelevance possibility by way of an example with a "reserve requirement": storage of k units of the consumption good from t to $t+1$ must be accompanied by storage of money from t to $t+1$ with value at t at least equal to ρk for $\rho\geqslant 0$. The physical environment of the example is $N(t)=1$ and $Y(t)=y>0$ for all t, $u^h(c_1, c_2)=u(c_1,c_2)$ with c_1 and c_2 being normal goods, and $x=(x_1)=\bar{x}>1$. Policy is given by $G(t)=0$, $K^g(t)=K^g$, $w_1^h(t)=y$ and $w_2^h(t)=K^g(\bar{x}-1)$ for all $t\geqslant 1$, and $\Sigma_h w_{2i}^h(0)=K^g(0)x(1)$. I will descirbe the dependence of the stationary equilibrium on the parameter K^g.

At any price s, profit from storing k units of the consumption good from t to $t+1$ consists of the profit from storing the good and the profit from storing the required money; namely, $(s\bar{x}-1)k+[sp(t+1)/p(t)-1]\rho k$. It follows that at $p(t+1)=p(t)=p>0$, $s\leqslant(1+\rho)/(\bar{x}+\rho)<1$ in any competitive equilibrium. It also follows that no additional money is stored at any $p(t+1)=p(t)=p>0$. That being so, the relevant version of (7) implies $c_2=(\bar{x}+\rho)K^p+K^g(\bar{x}-1)$. This and (1) imply $c_1=y-(1+\rho)K^p$. Then, letting $v(c_1, c_2)$ denote the function $u_1(c_1, c_2)/u_2(c_1, c_2)$, we may summarize (1)–(4) and (7) by

$$(18)\quad v[y-(1+\rho)K^p,(\bar{x}+\rho)K^p$$
$$+K^g(x-1)]=(\bar{x}+\rho)/(1+\rho)$$

if $K^p>0$.

A second condition on K^p and K^g (and p) is the relevant version of (5),

$$(19)\quad K^g=\rho K^p-p\overline{M}$$

where $\overline{M}$ is what the money supply would be if $K^g=0$.

Since $p\overline{M}\geqslant 0$, only values of K^g consistent with $K^g\leqslant\rho K^p$ and (18) give rise to a stationary equilibrium. Such values consist of the interval $[0,\tilde{K}^g)$, where $\tilde{K}^g$ is the solution for K^g to $K^g=\rho K^p$ and (18). (That $\tilde{K}^g>0$ and unique follows from the assumptions made about the utility function u.) For any K^g in $[0.\ \tilde{K}^g)$, the stationary equilibrium is found by solving (18) for K^p, and (19) for p. It is immediate that p is decreasing in K^g.

In this example, as in that of the last section, open-market operations have the usually asserted qualitative effects on the price of money. This implies that the welfare of the current old (at $t=1$) is affected in a similar way by such operations. But there the similarity ends. In Section V, $K^g>\overline{K}$ implies net taxes on the young and, in simple examples, makes everyone worse off than they are with $K^g\leqslant\overline{K}$. In this section, $K^g>0$ implies a net subsidy to the young and makes them better off than with $K^g=0$. In both cases, unchanged fiscal policy is not consistent with different government portfolios. Moreover, open-market operations seem to be consistent with "neutrality" in the sense of an unchanged real equilibrium only when the irrelevance proposition holds. When it holds, neutrality is accompanied by an unchanged price of money.

VII. Concluding Remarks

Most economists are aware of considerable evidence showing that the price level and the amount of money are closely related. That evidence, though, does not imply that the irrelevance proposition is inapplicable to actual economies. The irrelevance proposition applies to asset exchanges under some conditions. Most of the historical variation in money supplies has not come about by way of asset exchanges; gold discoveries, banking panics, and government deficits and surpluses account for much of it. Nothing in the models for which the irrelevance proposition holds denies that such occurrences alter the price level in the usual way.

Perhaps the main plea to be made for the irrelevance proposition is that it, and the environments in which it holds, should serve as the starting point for analyses of government asset exchanges. This is the same plea that is made for the Modigliani-Miller theory as a theory of corporate liability structures. The applicability of complete, competitive markets to open-market operations seems no more farfetched than its applicability to corporate liability structures. After all, economies of complete, competitive markets are ones in which a prohibition on the institution of limited liability does not matter.

Finally, a word of apology for the title of this paper is in order. The irrelevance proposition proved in Section III is defective because it leaves completely open the question: How broad is the class of environments for which the conclusion holds? Indeed, since the proposition is an arbitrage proposition, it may be possible to proceed without completely specifying the environment. For example, as regards individual choice, it is enough to establish that different government portfolios are consistent with the same budget set (equation (1)) for each individual. Moreover, although my notation uses the particular age composition assumed—two as opposed to n-period-lived people—and one good, none of this is necessary for the conclusion. Since I use all these extraneous assumptions, this paper ought to be viewed as providing only a suggestion for a general Modigliani-Miller theorem for open-market operations.

REFERENCE

J. E. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

Self-Selection in the Labor Market

By J. LUIS GUASCH AND ANDREW WEISS*

Often models of labor markets have assumed that firms know the productivity of all appplicants and pay wages proportionate to those productivities. In the presence of heterogeneity among the labor force and imperfect information by employers, this assumption is overly strong. A more common practice is for firms to offer a wage for a given job classification, and to test applicants to try to ensure a minimal level of performance.[1] The tests used by employers typically include a trial hiring period during which the applicant's performance is carefully monitored as well as perusual of the applicant's education record, previous job experience, and behavior during an interview.

Since tests are costly to administer, inaccurate, and imprecise, firms try to discourage applications from workers who do not meet their hiring standards. One way of discouraging the less qualified is to require applicants to pay a fee for being tested.[2] Imposing a cost for being tested discourages applications from individuals who believe their probability of passing the examination is low (as well as from poorer individuals if the marginal utility of income is decreasing). The use of an application fee has the effect of converting a one-part test into a two-part test; only workers who both perceive their probability of passing the test to be high and who actually pass the test are hired.

For example, if an apprenticeship program is viewed as a test, then below-market wages during the apprenticeship discourage applications from workers who would be less likely to successfully complete their apprenticeships (those workers who successfully finish the apprenticeship program will receive an increase in their wages). The difference between an applicant's wage in the training program and the wage he could obtain elsewhere is the fee for being tested.[3] This interpretation of apprenticeship implies that wages increase with job tenure, nòt because of the acquisition of human capital, but rather as a consequence of the combination of tests and wages to sort workers. The debate here is similar to that over sorting versus human capital theories of education; however, by focusing on sorting versus human capital explanations of low-wage apprenticeships, we can more readily generate testable hypotheses which distinguish between the two models. For example, the human capital approach predicts that productivity per manhour increases with job tenure, while a pure sorting model predicts that, for individual workers, their productivity is independent of their job tenure.[4]

*University of California-San Diego and Bell Laboratories, respectively. We benefited from illuminating conversations with Joseph Stiglitz. We also wish to thank Peter Fishburn, Bruce Greenwald, Henry Landau, Ed Zajac, and George Borts for helpful comments. The beginning of this research program was supported by a Sloan Foundation grant for Weiss. Work was continued while Guasch was a Research Fellow at CORE, Universite Catholique de Louvain, Belgium.

[1] Issues of productivity differences among workers paid a uniform wage has been discussed by Melvin Reder (1955, 1969), William Brown, and Weiss (1976), among others.

[2] One potential difficulty with this test-cum-fee strategy is that firms may have an incentive to take the fees and announce that workers have failed, without testing them, i.e., "take the money and run." In our model this problem can be ignored since firms, in general, are making positive profits. If they renege on their explicit contract to test applicants and hire those who pass, they would not attract future applicants and would find the expected value of their stream of profits to have fallen. In cases where reneging on contracts is profitable, we would expect legal instruments to arise to enforce contracts.

[3] Although, in principle, the screening mechanisms available to firms are numerous, we will focus specifically on the use of testing strategies, with or without application fees. The nontesting strategies, where the firm uses the wage offered as a means of maximizing labor input per dollar, have been analyzed in our 1980b article and by Weiss (1980).

[4] Of course, measurements of the average productivity of workers may show increases with job tenure if the

We develop a partial equilibrium model of firm behavior to analyze the conditions under which a combination test-fee is more profitable to the firm than either testing without a fee or hiring untested workers, and characterize the wage schedule which results when the test-fee combination is used to sort workers. In the Appendix, we explicitly consider the case where the "fee" takes the form of a low-wage apprenticeship program.

I. The Model

Consider a firm selecting the optimal hiring strategy when faced by a heterogeneous labor force. Each type or group of workers is characterized by an expected labor input (productivity), current earnings (acceptance wage), nonlabor endowment, and von Neumann-Morgenstern preferences over lotteries in earnings. The firm is characterized by a production function and there is only one test available to screen applicants. Throughout the text we will use the following notation:

$g(\cdot)$ = the production function which is assumed to be continuously differentiable and concave

A = the cost of the test per applicant tested

L = the number of workers tested by the firm

c = the fee charged by the firm to take the test

p_i = the probability of a type i worker passing the test

Q_i = the expected labor input of type i workers

w_i = the present value of the acceptance wage of a type i worker (referred to as the acceptance wage)

w^* = the present value of the wage offered by the firm to all workers passing the test (referred to as the acceptance wage).

Note that each firm offers a single wage to successful applicants.

To simplify the analysis we assume that there are only two types of workers 1 and 2, α is the percentage of type 1 workers in the labor force, and $Q_1 > Q_2$, $w_1 > w_2$, and $p_1 > p_2$. Also it is assumed that workers who are not employed by the firm are employed elsewhere.[5] A strategy is a wage w, and a decision to (or not to) administer a test to applicants, with or without a fee c. The set of strategies can be partitioned as

S_I: offering a wage w, a fee c, and a test to applicants such that only type 1 workers apply to work (self-selection), clearly $w \geqslant w_1$;

S_{II}: offering wage $w \geqslant w_1$ and testing all applicants, without imposing an application fee;

S_{III}: offering a wage between w_1 and w_2 and hiring only type 2 applicants without testing;

S_{IV}: offering a wage $w \geqslant w_1$ and hiring randomly among applicants without testing.

The firm will choose the strategy that renders the largest expected profit. A strategy is a means of purchasing an input and, if the labor supply constraint is not binding (an assumption we make throughout this paper), the profit-maximizing strategy is the one in which the unit cost of labor is lowest. For a proof of this assertion, see Weiss (1980) or our 1980b article.

The profitability of each strategy is determined by the degree of risk aversion of type 1 and type 2 workers, differences in their productivity-acceptance wage ratio, misperceptions of the probability of passing the test, and differences in their nonlabor endowments. Our aim is to investigate how these characteristics affect the feasibility and profitability of the self-selection strategy. To single out the effects of each factor we analyze them separately.

workers who fail the test are less able than those who pass the test. Weiss and B. Greenwald are conducting an empirical investigation to test the relation between tenure and productivity for semiskilled workers in a number of manufacturing plants. An analysis of data from one of these plants shows that the productivity of individual workers does not rise with tenure after their third month on the job.

[5] The justification for $w_1 > w_2$ relies on the assumption that each worker is paid a wage correlated with his expected marginal product in that alternate job.

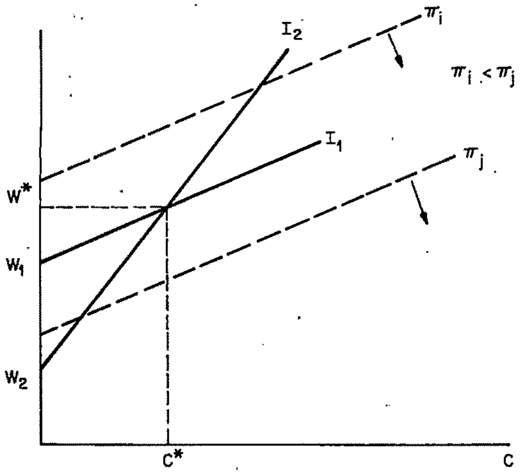


FIGURE 1

Under a successful self-selection scheme, only type 1 workers apply to work for the firm. Let I_i denote the set of wages and positive fees which yield the same expected utility to workers i as a certain wage of w_i. If type 1 workers are risk neutral and accurately informed about their probabilities of being accepted, they are indifferent among all the combinations of w and c which yield them an expected income of w_1. The firm is similarly indifferent among these possibilities since the slope of the iso-profit line is the same as the slope of the indifference curve of type 1 worker between wages and fees, I_1.[6] Referring to Figure 1, the constant utility line of type 1 workers, I_1, is flatter than the constant utility line of type 2 workers, I_2. The slope of the former is $1/p_1$, the slope of the latter is $1/p_2$. Thus, for any value of w and c lying along I_1 and to the right of (c^*, w^*), self-selection will occur. Since the profits of the firm are the same at any point along I_1, changes in w_2 or in the slope of I_2 do not affect the firm's profits. On the other hand, when workers are risk averse, wages and fees that lie on the indifference curve of a type 1 worker all yield different profits. This case is discussed in Section III.

II. Screening Strategies with Risk-Neutral Workers

Under risk neutrality, nonlabor endowments are irrelevant, so we will not consider them.

PROPOSITION 1: *If workers are risk neutral and have no misperceptions about their probability of passing the test, then*

a) *A self-selection strategy exists.*[7]

b) *There exists a self-selection strategy which dominates all testing strategies without self-selection.*

c) *There exists a self-selection strategy which dominates any strategy in S_{III} (hiring type 2 alone) iff*

$$\frac{w_2}{Q_2} - \frac{w_1}{Q_1} > \frac{A}{p_1 Q_1}$$

d) *There exists a self-selection strategy which dominates any strategy in S_{IV} (hiring randomly at wage w_I) iff*

$$\frac{A + w_1 p_1}{p_1 Q_1} < \frac{w_1}{aQ_1 + (1-a)Q_2}$$

PROOF:

Since $p_1 > p_2$, $p_1/p_2 > 1$; note also that the function $f(w) = (w - w_2)/(w - w_1)$ approaches 1 as w goes to ∞. Therefore, for any $k = p_1/p_2 > 1$ there is a w^* such that $\forall$ $w > w^*$, $f(w) < k$. Equivalently, we can write

[6] Actually, the profit lines depicted in the figures are to be understood as being active only on the self-selection region of the (w, c) space. The isoprofit lines gives the profit generated by the pair (w, c), contingent on self-selection taking place. But clearly, for any (w, c) outside of the region where self-selection occurs those isoprofit lines are not well defined.

[7] The existence of a self-selection strategy depends upon an underlying assumption of risk neutrality: that workers can pay fees of any magnitude. An alternative specification is plausible: if workers are not permitted to contract to pay fees that exceed the present value of their present lifetime earnings, then for any $c > w_2$, no type 2 workers would apply; clearly these fees will exclude type 2 workers. On the other hand, we could permit bankruptcy and allow workers to sign contracts which cannot be fulfilled. In this case workers would be indifferent between payoffs of zero and negative payoffs. In that regime it may not be possible to derive a combination of wages and application fees which partition the labor force.

that for all $w>w^*$, $p_1(w-w_1)>p_2(w-w_2)$. Since the individuals are risk neutral, to exclude type 2 workers it will suffice to charge a fee c such that $p_1(w-w_1)>c>p_2(w-w_2)$. To show (b) it suffices to prove that for every $\hat{s}\in S_{II}, \exists s^*\in S_I$ such that $\pi_I(s^*)>\pi_{II}(\hat{s})$. The profits of the firm inducing self-selection are $\pi_I(s^*)=g(p_1Q_1L)-w^*p_1L+c^*L-AL$. If the firm is maximizing profits, then $w^*=c^*/p_1+w_1$, and max $\pi_I(s^*)=g(LP_1Q_1)-w_1p_1L-AL$. Note that we can write the minimum cost per efficiency unit of labor as $E^*=(A+w_1p_1)/p_1Q_1$. On the other hand, a firm that does not charge a fee to screen workers but that does test applicants has a profit function

$$\pi_{II}(\hat{s})=g\big(\alpha p_1Q_1L+(1-\alpha)p_2Q_2L\big)$$
$$-w\big(\alpha p_1+(1-\alpha)p_2\big)L-AL$$

where $w\geqslant w_1$ and its minimum cost per efficiency unit of labor is

$$\hat{E}=\frac{A+w_1\big(\alpha p_1+(1-\alpha)p_2\big)}{\alpha p_1Q_1+(1-\alpha)p_2Q_2}$$

To show that a strategy s^* yields larger profits than strategy $\hat{s}$, we need only to prove that $E^*<\hat{E}$. Noting that $p_1>\alpha p_1+(1-\alpha)p_2$, we can write the following inequalities:

$$E^*=\frac{A+w_1p_1}{p_1Q_1}<\frac{A+w_1[\alpha p_1+(1-\alpha)p_2]}{Q_1[\alpha p_1+(1-\alpha)p_2]}$$
$$<\frac{A+w_1[\alpha p_1+(1-\alpha)p_2]}{\alpha p_1Q_1+(1-\alpha)p_2Q_2}=\hat{E}$$

To show (c), let $\tilde{E}$ represent the minimum cost per efficiency unit of labor for a firm pursuing a strategy $\tilde{s}\in S_{III}$, not testing workers. Then $\tilde{E}=w_2/Q_2$ and $E^*<\tilde{E}$ iff $A/p_1Q_1<w_2/Q_2-w_1/Q_1$. Similarly, the proof of (d) follows directly from the cost per efficiency unit of labor of S_I and S_{IV}, respectively.

Note that, although all the applicants who take the test are of type 1, the firm hires only the type 1's who pass the test in order to deter the type 2's from applying. This rejection of type 1 failures persists even if the firm ”knows” that it is rejecting qualified workers.

By assuming that p_1 and p_2 are exogenous parameters, we have thus far precluded randomized testing. As Joseph Stiglitz (1975a) has pointed out, allowing randomization can lead to open-set problems in sorting models. In the context of our model, Stiglitz's objection can be phrased as follows: a profit-maximizing firm which is testing workers may exempt a fraction $1-\gamma$ of its applicants from the test. By increasing the wage and fee it offers to applicants, the firm can ensure itself that all the untested applicants will be of type 1 and that those workers receive an expected income w_1. Therefore, the quality of its applicant pool is unaffected by the randomizing strategy. But, by hiring the untested workers, the firm is able to reduce its testing costs and decrease its cost per efficiency unit of labor. Since

$$E^*=\frac{\gamma A+(1-\gamma)w_1+\gamma p_1w_1}{(1-\gamma)Q_1+\gamma p_1Q_1}$$
$$=\frac{w_1}{Q_1}+\frac{\gamma A}{(1-\gamma)Q_1+\gamma p_1Q_1}$$

a decrease in γ with the untested workers being hired will decrease E^*. The open-set problem arises because profits are monotonically increasing in $1-\gamma$ until $\gamma=0$, at which point the firm is hiring without a test. This may yield lower profits than some strategy in S_I.

We can avoid this dilemma and put more realism into the model presented above by dropping the assumption that the workers who pass the test are identical to those who fail the test. If those type 1 workers who pass the test are more productive than those who fail, a firm may wish to test all its applicants. We will denote the ability of type 1 workers who pass the test as Q_1^s and those who fail the test as Q_1^f. Then the cost per efficiency unit of labor is

$$E^{**}=\frac{\gamma A+(1-\gamma)w+\gamma p_1w-c}{(1-\gamma)\big[p_1Q_1^s+(1-p_1)Q_1^f\big]+\gamma p_1Q_1^s}$$

Through some fairly straightforward calculations we find that randomization increases profits iff $(Q_1^s - Q_1)p_1 w_1 < AQ_1$. As we would expect, increases in the cost of the test or decreases in the difference between type 1 workers who pass and the average type 1 worker make firms more likely to randomize. Similarly, increases in p_1 makes randomized testing less likely. The inclusion of w_1 serves simply as a scale factor. One of the interesting implications of this result is that, for risk-neutral workers, a firm either does not randomize at all or tries to completely randomize, confronting the open-set problem.

The above results hinge on the assumption that workers correctly perceive their probability of passing the test. If perceptions are incorrect, employing an application fee to screen workers may no longer dominate the testing program without an application fee. To see that, let p_1' and p_2' be the perceived passing probabilities of type 1 and 2 workers, respectively, and assume that $p_1' \leqslant p_1$ and $p_2' \geqslant p_2$. Then, as the difference between the true and perceived probabilities of passing the test increases, self-selection becomes a less-profitable strategy (with $p_1' > p_2'$, otherwise a self-selection strategy will not exist); as the perceived probabilities of passing the test by the two groups approach one another, testing without an application fee dominates the self-selection strategy. This result follows from noting that the cost of labor per efficiency unit is $E=$

$$\frac{p_1[p_1'w_1 - p_2'w_2] - p_1'p_2'[w_1 - w_2] + A[p_1' - p_2']}{p_1 Q_1^s(p_1' - p_2')}$$

And as $p_1' \rightarrow p_2'$, $E \rightarrow \infty$.

In general, we might expect some uncertainty among workers as to the group they belong to, so that type 1 workers would underestimate their probability of passing the test, while type 2 workers would overestimate their passing probability. Under those circumstances better information by either group concerning their true probability of passing the exam increases the profitability of the self-selection strategy.

III. Screening Strategies with Risk-Averse Workers

We assume here that workers are risk averse, have accurate perceptions of their probability of passing the test, and have no nonlabor endowments. They possess von Neumann-Morgenstern utility functions over wealth $U(x)$ where $U(x) \rightarrow -\infty$ as $x \rightarrow 0$, $U(x) \rightarrow \infty$ as $x \rightarrow \infty$, $U'(x) > 0$, and $U''(\cdot) < 0$.[8] Although we assume that the preferences of both 1 and 2 workers are represented by the same utility function, this assumption is not essential to the results derived below. Because the profitability of any strategy $s \notin S_1$ does not change with risk aversion, we confine our analysis to the effects of risk on the self-selection strategies.

The constant utility wages and fees or indifference curves of types 1 and 2 workers are

$$I_1 = \{(w,c) | U(w_1) = p_1 U(w-c) + (1-p_1)U(w_1 - c), w \geqslant 0, c \geqslant 0\}$$

$$I_2 = \{(w,c) | U(w_2) = p_2 U(w-c) + (1-p_2)U(w_2 - c), w \geqslant 0, c \geqslant 0\}$$

Notice that the indifference curves of risk-averse workers are not straight lines, hence the values of (w,c) lying on I_1 do not all yield the same profits to the firm (which was the case when the workers were risk neutral). Let us denote by B the set of w and c such that for any $(w,c) \in B$, type 1 workers will apply and type 2 workers will not apply:

$$B = \{(w,c) | U(w_1) \leqslant p_1 U(w-c) + (1-p_1)U(w_1 - c)$$

and

$$U(w_2) \geqslant p_2 U(w-c) + (1-p_2)U(w_2 - c)\}$$

[8] None of the qualitative results derived below would be affected if, instead of assuming $\lim_{x \rightarrow 0} U(x) = -\infty$ and $\lim_{x \rightarrow \infty} U(x) = \infty$, we had assumed that a worker of type i could not contract to pay a fee $c \geqslant w_i$. In that case, the point on a type 1 worker's constant utility curve where $c = w_2$ is a separating contract.

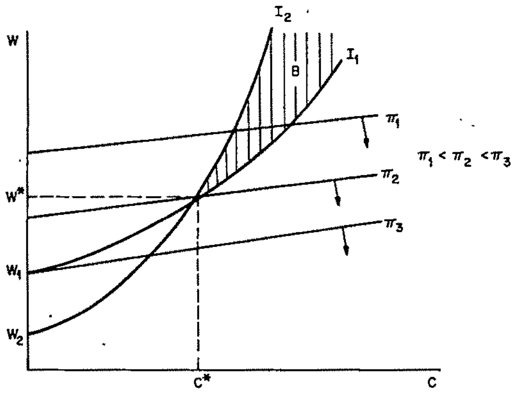


FIGURE 2

On Figure 2 we represent the indifference curves I_1 and I_2 as well as the isoprofit lines π_i. The set B is not empty. Since $\lim_{x\rightarrow\infty}U(x)=\infty$ and $\lim_{x\rightarrow 0}U(x)=-\infty$, the value of w satisfying I_2 approaches ∞ as $c\rightarrow w_2$, while a finite value of w satisfies I_1 when $c=w_2$. This ensures that there exist sufficiently large fees such that I_2 crosses I_1 at $c<w_2$. (The reader should note that risk aversion alone is not sufficient to ensure that I_1 and I_2 cross at $c<w_2$. For example if $U(x)\rightarrow 0$ as $x\rightarrow 0$ then if $U(w_1)>U(w_2)/p_2, I_1$ will not cross I_2 at $c<w_2$. On the other hand, if we do not allow type 2 workers to contract to pay fees greater than w_2, then $w_2<c<w_1$ will always separate type 1 from type 2 workers.) We now derive the element of B which maximizes the firm's profits.

PROPOSITION 2: *Let* $(w^*,c^*)\in I_1\cap I_2$, *then* $\pi(w^*,c^*)>\pi(w,c)$ *for any* $(w,c)\in B$ *and* $(w,c)\notin I_1\cap I_2$.

PROOF:

The profit function for the self-selection strategy is

$$\pi(w,c)=g(Lp_1Q_1)-wp_1L$$
$$+cL-AL$$

thus the iso-profit lines on the space (w,c) are straight lines with slope $1/p_1$. If we denote by $t_1(w,c)$ and $t_2(w,c)$ the slope of the indifference curves of 1 and 2, respectively, then

$$t_1(w,c)=1+\frac{(1-p_1)U'(w_1-c)}{p_1U'(w-c)}$$

and $$t_2(w,c)=1+\frac{(1-p_2)U'(w_2-c)}{p_2U'(w-c)}$$

with $t_i(w_i,0)=1/p_i$. Furthermore, by the concavity of U, I_1 is convex, as is I_2. Thus, at any point except at $(w_1,0)$, the slope of the indifference curve I_1 is greater than the slope of the iso-profit line, and the lower the intersect of the iso-profit line with the vertical axis, the higher the profits. When B is nonempty, the curves I_1 and I_2 intersect. Given the assumption of strict concavity on the utility function, the intersection of the indifference curves I_1 and I_2 will be a single point denoted by (w^*,c^*).[9] This combination generates the maximum profit among self-selection pairs (w,c); i.e., $\pi(w^*,c^*)>\pi(w,c)$, for $(w,c)\in B$ and $(w^*,c^*)\neq(w,c)$.

Since the intersection of I_1 and I_2 can occur at any value of (w^*,c^*), subject to $w^*>w_1$ and $c^*>0$, there is no lower bound on the maximum profits generated by a self-selection strategy. On the other hand, risk aversion does not affect the profitability of any other strategy $s\notin S_I$, therefore, there are values of w_1, w_2, p_1, p_2 and $U(\cdot)$ for which testing without self-selection can earn higher profits than testing with a self-selection mechanism. This result differs from the result in Proposition 2 in which we found that for *risk-neutral* workers a self-selection strategy always dominates testing without self-selection.

Given that I_1 will always be above the isoprofit line through $(w_1,0)$, which gives the profitability of the self-selection strategy under risk neutrality, we can state

PROPOSITION 3: *Risk aversion decreases the profitability of self-selection strategies.*

[9] We will take the limit point (w^*,c^*) as the choice of the maximum-profit strategy under risk aversion as if this pair would also work; the order of magnitude of the error in the computation of profits can be made as small as desired.

It follows, that under risk aversion, any variation in the parameters that produces an increase in the curvature of I_1 or an upward shift of I_1, such as an increase in w_1 or a decrease in p_1, decreases the profitability of the self-selection strategy. On the other hand, increases in w_2 or decreases in p_2 increase the profitability of the self-selection strategy. Note that if type 1 workers were risk neutral, but type 2 workers were risk averse, the relative profitability of different strategies would be the same as when both types of workers are risk neutral.

There is a clear intuition behind these results. Similarities in probabilities of passing the test make the self-selection procedure more difficult (increasing the minimum fee at which self-selection occurs). Because of the risk aversion of type 1 workers, upward shifts in the optimum value of (w, c) separating the two groups decreases the profit of the firm employing self-selection. As in the risk neutral case, an increase in testing costs increases the profitability of self-selection strategies relative to testing without self-selection.

The open-set problem that always appears with randomization under risk neutrality when $Q_1^s = Q_1$ may not arise when workers are risk averse. On the other hand, risk aversion alone (even if $\lim_{x\to 0} U(x) = -\infty$) is not sufficient to ensure that the open-set problem is avoided.

To see this, again let γ denote the fraction of applicants tested, and let the constant utility wages and fees for types 1 and 2 be represented by

$$U(w_1) = (1-\gamma(1-p_1))U(w-c) + \gamma(1-p_1)U(w_1-c)$$

and

$$U(w_2) = (1-\gamma(1-p_2))U(w-c) + \gamma(1-p_2)U(w_2-c)$$

The cost of labor per efficiency unit can be written as

$$E = \frac{\gamma A + w(\gamma p_1 + 1 - \gamma) - c}{\gamma p_1 Q_1^s + (1-\gamma) Q_1}$$

clearly, the open set-problem is avoided if $\lim_{\gamma\to 0}[dE/d\gamma] < 0$. Collecting terms,

$$sign\left[\lim_{\gamma\to 0}\frac{dE}{d\gamma}\right] = sign\left[AQ_1 - wp_1(Q_1^s - Q_1) + c(p_1Q_1^s - Q_1) + Q_1\lim_{\gamma\to 0}\frac{dw}{d\gamma} - Q_1\lim_{\gamma\to 0}\frac{dc}{d\gamma}\right]$$

Given that A is finite and that w goes to $w_1 + w_2$ and c goes to w_2 as $\gamma \to 0$, $dE/d\gamma$ will surely be negative if $\lim_{\gamma\to 0}[dw/d\gamma - dc/d\gamma] = -\infty$. Using Cramer's rule, and recalling that $\lim_{x\to 0} U(x) = -\infty$ and $\lim_{x\to\infty} U(x) = \infty$, we find

$$\lim_{\gamma\to 0}\left[\frac{dw}{d\gamma} - \frac{dc}{d\gamma}\right] = -\infty \quad \text{iff} \quad \lim_{x\to 0}\left[\frac{U(x)}{U'(x)}\right] = -\infty$$

This condition holds for some but not all concave utility functions satisfying $\lim_{x\to 0} U(x) = -\infty$ (we can ignore the restriction of $\lim_{x\to\infty} U(x) = \infty$ by allowing piecewise functions). For example if $U(x) = log\, x$ then $\lim_{x\to 0} U(x)/U'(x) = -\infty$ and firms will never wish to randomize completely. On the other hand if $U(x)$ behaves as $-x^{-1}$ in the neighborhood of $x = 0$, then $\lim_{x\to 0} U(x)/U'(x) = 0$ and

$$\lim_{\gamma\to 0}\left[\frac{dw}{d\gamma} - \frac{dc}{d\gamma}\right] = \frac{(1-p_1)[U(w_1) - U(w_1 - w_2)]}{U'(w_1)}$$

Substituting into $\lim_{\gamma\to 0} dE/d\gamma$ we find that in this case risk aversion avoids the open-set problem only if

$$AQ_1 - p_1w_1[Q_1^s - Q_1] + (1-p_1)Q_1\left[\frac{U(w_1) - U(w_1 - w_2)}{U'(w_1)} - w_2\right] < 0$$

Although these are sufficient, but not necessary, conditions for firms to choose an interior value of γ, one can easily show that risk aversion will not always lead to an interior value of γ.

$$sign\left[\frac{dE}{d\gamma}\right]=sign\left\{AQ_1-wp_1[Q_1^s-Q_1]\right.$$
$$+c[p_1Q_1^s-Q_1]+[\gamma p_1Q_1^s+(1-\gamma)Q_1]$$
$$\left.\times\left[(\gamma p_1+1-\gamma)\frac{dw}{d\gamma}-\frac{dc}{d\gamma}\right]\right\}$$

From the requirement that the constant utility curves of type 1 and type 2 workers cross, we can show that for $\gamma>0$ all the terms of $dE/d\gamma$ are finite. If A is sufficiently large and $\lim_{x\to 0}U(x)/U'(x)$ is finite, $dE/d\gamma$ will be positive for all values of $\gamma\geqslant 0$ and firms will face an open-set problem when choosing the proportion of workers to test.

IV. The Effect of Wealth Differences Upon Self-Selection

Thus far we have ignored differences in nonlabor endowments. Let us introduce an additional type of worker denoted by $2R$. Type $2R$ workers have the same alternative wage as do type 2 workers and the same probability of passing an examination but have a nonlabor endowment k. We also assume that type 2 and type $2R$ workers have the same utility function, and that the utility function is characterized by decreasing absolute risk aversion, $d/dx(-U''(x)/U'(x))<0$.

The indifference curve for a type $2R$ worker, I_{2R}, intersects the vertical axis at w_2 and has the same slope at that point of $1/p_2$; however, since the slope of type $2R$'s indifference curve is smaller than type 2's, I_{2R} is always below I_2 except at the original point $(w_1,0)$ where they coincide. This difference in slopes follows directly from decreasing absolute risk aversion, which implies

$$1+\frac{(1-p_2)}{p_2}\frac{U'(w_2-c)}{U'(w-c)}$$
$$>1+\frac{(1-p_2)U'(w_2-c+k)}{p_2U'(w-c+k)}$$

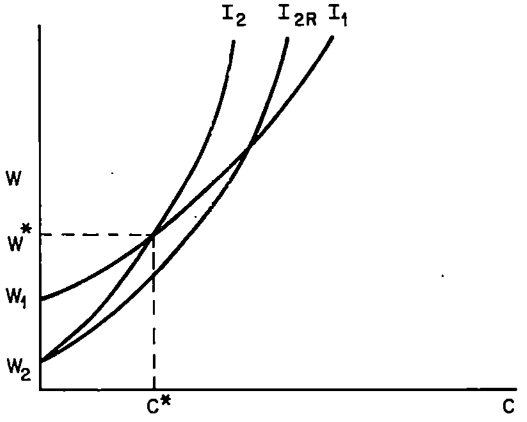


FIGURE 3

Thus a self-selection mechanism offering the wage fee combination of (w^*,c^*) illustrated in Figure 3 will not exclude type $2R$ workers. Any firm choosing a self-selection procedure has two choices to make. It can either exclude only type 2 workers or it can exclude both type 2 and type $2R$. If there are relatively few type $2R$ workers and their unearned income is relatively high, firm profits are maximized by only excluding type 2 workers.

In fact, it may not be possible to hire type 1 workers and exclude type $2R$ workers. In the extreme case in which $p_1=p_2$ and where the wealth of type $2R$ workers yields them an unearned income greater than w_1-w_2 then I_2 will never intersect I_1 and there doesn't exist any wage-fee combination which will exclude the low-productivity type $2R$ workers. To the extent that self-selection mechanisms favor wealthy individuals, one might question the value of government regulations which increase testing costs. As we have shown, such measures tend to increase the use of self-selection mechanisms, and thus favor wealthy applicants.

V. Conclusion

Since both the wage structure and organization of firms may be explained (at least in part) by the use of self-selection strategies, it is important to identify the theoretical conditions under which those strategies are feasible and profitable. As the preceding sentence suggests, there are two aspects to this question. First, is there always a wage-fee combination which effectively divides the labor force? Second, is there an effective self-selection strategy which is more profitable than testing without a fee or hiring without any testing?

We have shown that if risk neutral or risk averse workers have accurate information about their probabilities of passing the test, then there is always a wage-fee combination which only attracts type 1 workers.

In an economy where workers correctly perceive their true probability of passing the examination and are risk neutral, a self-selection mechanism that partitions the labor force will increase the profits of a firm that tests its workers. However, if workers are risk averse, it is possible that there is no self-selection strategy that dominates testing workers without self-selection. As the acceptance wages of the two groups of workers become further apart, and as their probability of passing the test becomes closer together, the profitability of the self-selection strategy falls relative to testing without self-selection. For both risk-neutral and risk-averse workers, if the more able workers underestimate their probability of passing the examination, the profits of a firm that does employ a self-selection mechanism will fall. The reason is that the firm pays a wage-fee corresponding to the perceived, not true, pass probability of the high-ability workers.

For all these reasons, it is not surprising that we do not see self-selection procedures always used when workers are hired. On the other hand, in cases where individuals have significantly more accurate priors concerning their probability of passing the test than do firms, we do see the utilization of self-selection devices.

The natural extensions of this model are to allow for an arbitrary distribution of many different types of workers, to allow acceptance wages to be uncorrelated with expected labor inputs, to introduce competition among firms, and to let the passing level of the test be a choice variable of the firm. These extensions are undertaken in our 1980a,c papers, where the analysis is restricted to risk-neutral workers with accurate perceptions of passing each test. Those papers present necessary and sufficient conditions for the existence of a Nash equilibrium with many firms and free entry. Three characteristics of this equilibrium, when the distribution of productivities within types is Gaussian, are that workers who pass the test receive a wage net of testing fees above their expected marginal product, workers who fail receive a net wage below their expected marginal product, and the application fee always exceeds the true cost of the test.[10] We conjecture that these same results hold for the more general case of risk averse workers with wealth differences.

Appendix: Existence of a Separating Apprenticeship Program

Although the analysis of this paper has focused upon workers paying an explicit application fee, much of the discussion and interpretation has concerned apprenticeship programs as tests. If an apprenticeship program is used as the self-selection mechanism, then the test is more costly for the more able type 1 workers than for the less able type 2 workers (the former have larger foregone incomes). Below we derive the conditions under which an apprenticeship program can serve as a self-selection device.

We assume that workers are not paid during the apprenticeship program, and that utility is defined only over lifetime earnings, that is, workers are indifferent between different time paths each of which generate the same lifetime earnings.

We will denote by $(1-\beta)$ the fraction of an employee's work life which the training program occupies. Thus type 1 workers will participate in an apprenticeship program iff

$$\text{(A1)} \quad \beta\left(p_1 U(w^*)+[1-p_1]U(w_1)\right) \geqslant U(w_1)$$

[10] The intuition is that the zero-profit condition determines the expected wage each type of worker receives. The workers are sorted by increasing the penalty to workers failing the test. This penalty necessitates that each failure receives a net wage below his net marginal product while the zero-profit condition implies that those workers who pass the test receive net wages above their net marginal product.

Type 2 workers will be deterred from participating iff

$$\text{(A2)}\quad \beta\big(p_2U(w^*)+[1-p_2]U(w_2)\big)\leqslant U(w_2)$$

Therefore, a self-selection strategy that attracts only type 1 applicants will exist if there exist values of β and w^* such that (A1) and (A2) both hold. We can rewrite (1) and (A2) as

$$\text{(A1}'\text{)}\quad \beta U(w^*)\geqslant\frac{U(w_1)-[1-p_1]U(w_1)\beta}{p_1}$$

$$\text{(A2}'\text{)}\quad \beta U(w^*)\leqslant\frac{U(w_2)-[1-p_2]U(w_2)\beta}{p_2}$$

A self-selection strategy exists iff

$$\text{(A3)}\quad \frac{U(w_2)-[1-p_2]U(w_2)\beta}{p_2}\geqslant\frac{U(w_1)-[1-p_1]U(w_1)\beta}{p_1}$$

or equivalently iff

$$\text{(A4)}\qquad \frac{p_1}{p_2}\geqslant\frac{U(w_1)}{U(w_2)}$$

Equation (A4) has a clear intuitive explanation. As p_1 and p_2 become closer together, the expected benefit to type 1 approaches the expected benefit to type 2 workers, and thus it becomes more difficult to separate the two types. On the other hand, as $U(w_1)$ becomes closer to $U(w_2)$, the cost of the apprenticeship program for the type 1 group approaches the cost for type 2, and thus it becomes easier to only attract type 1 workers.

REFERENCES

G. Akerlof, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart. J. Econ.*, Aug. 1970, *89*, 488–500.

William Brown, *Piecework Abandoned*, London: Heinemann Press 1962.

J. L. Guasch and A. Weiss, (1980a) "Wages as Sorting Mechanisms in Competitive Markets with Asymmetric Information: A Theory of Testing," *Rev. Econ. Stud.* July 1980, *47*, 653–64.

______**and**______, (1980b) "Adverse Selection by Markets and the Advantage of Being Late," *Quart. J. Econ.*, May 1980, *94*, 453–66.

______**and**______, (1980c) "Equilibrium Wage Distributions with Endogenous Hiring Standards," disc. paper 80–16, Univ. California-San Diego, 1980.

M. W. Reder, "The Theory of Occupational Wage Differentials," *Amer. Econ. Rev.*, Dec. 1955, *45*, 833–52.

______, "Wage Structure and Structural Unemployment," *Rev. Econ. Stud.*, Oct. 1969, *31*, 309–22.

M. Rothschild and J. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quart. J. Econ.*, Nov. 1976, *90*, 629–66.

J. K. Salop and S. C. Salop, "Self-Selection and Turnover in the Labor Market," *Quart. J. Econ.*, Nov. 1976, *90*, 619–28.

S. C. Salop, "Wage Differentials in a Dynamic Theory of the Firm," *J. Econ. Theory*, Aug. 1973, *6*, 321–44.

A. Michael Spence, (1974a) *Market Signalling: Information Transfer in Hiring and Related Processes*, Cambridge, Mass.: Harvard University Press 1974.

______, (1974b) "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *J. Econ. Theory*, Mar. 1974, *7*, 196–332.

J. E. Stiglitz, (1975a) "The Theory of Screening, Education, and Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

______, (1975b) "Information and Economic Analysis," in Michael Parkin and A. R. Nobay, eds., *Current Economic Problems*, Cambridge: Cambridge University Press 1975, 27–52.

A. Weiss, "A Theory of Limited Labor Markets," unpublished doctoral dissertation, Stanford Univ. 1976.

______, "Job Queues in Labor Markets with Flexible Wages," *J. Polit, Econ.*, June 1980, *88*, 526–39.

The Effect of Changes in the Population on Several Measures of Income Distribution

By SAMUEL A. MORLEY*

My reading of the distribution literature convinces me that the effect of population growth on the interpretation of the statistics being used to measure distributional performance has gone largely unappreciated. Commonly, distribution is measured by a comparison of income shares or real income growth rates over time of groups such as the poor or the rich. Yet, under population growth there is an important distinction to be made between the "poor" or the "rich," and what I call the "base-period poor or rich." The former is the group at the top or bottom of the distribution, the latter is the particular group of people who were poor or rich in the base period. The shares and income growth rates of the poor or rich reported in all studies refer to the poor or rich in the first sense, not the second. But most people who use those statistics seem to think they measure what has happened to the base period, poor or rich. They do not. Calculations for Brazil show that the difference between income growth for the poor and the base-period poor are large and significant. I do not claim that the standard measures are wrong, but rather that they do not tell us what happened to the income distribution of a base-period population, or whether the base-period poor shared the benefits of income growth over time.[1]

In this paper I would like to explore the relationship between population growth and two of the most widely used distribution statistics—the share of the rich and the poor in total income and the rate of growth of real income of the rich and the poor. In Section I, I present a simple model of the size distribution of income and use it to show the behavior of income shares and real income over time when the labor force is growing. As shall be seen, both measures are affected by the amount of labor force growth and by where new entrants came into the income pyramid. In Section II, I apply the theoretical model to Brazil and show that there are large differences between the income growth rates of the poor or rich, and the base-period poor or rich. Furthermore the behavior of income shares at the bottom, which has occassioned so much adverse comment about inequitable growth, is shown to be largely a result of population growth and a change in the profile of lifetime income, rather than relatively low rates of growth of income of the poor as has been widely reported.

I have chosen Brazil as an example of the difficulty of interpreting the standard inequality measures in the face of rapid population growth. It is a good country to use for this purpose for it is generally held up as the ultimate example of inequitable growth. Over the 1960's, Brazil's Gini coefficient rose from .50 to .57, its Lorenz curve shifted wholly to the right and the average per capita real income of the bottom 40 percent rose by only 18 percent compared to an average of 37 percent for the entire labor force and 67 percent for the top decile. Most previous work on distribution in Brazil has focused on the causes of rising inequality (see Edmar Bacha and Lance Taylor; Albert Fishlow; Carlos Langoni; Morley; Morley and Jeffrey Williamson; Ricardo Tolipan and Arthur Tinelli, and John Wells). I do not wish to add to this literature. Rather, by showing the large effect of population growth on the in-

*Vanderbilt University. I am indebted to Alan Blinder, Sheldon Danziger, Keith Horner, Graham Pyatt, George Sweeney, and a referee for their comments on a previous draft of this paper.

[1] Gary Fields fell into exactly this confusion. He argued that we should measure distribution performance by the absolute income gains for the poor. But he didn't distinguish between the poor and the base-period poor. Hence, what he measures as income gains by the poor is not the gains of 1960's poor, even though he writes as if it were. See particularly p. 580.

terpretation of inequality measures in the Brazilian case, I hope to convince the reader that income shares or growth rates cannot without adjustment tell us much about the degree of progressivity of a particular country's growth strategy.

I. The Formal Model

I seek here a formal way of investigating the relationship between population growth and several of the common distribution statistics.[2] To help do that, let us order the members of a base-year population by income level, and assign to each person an index number, X_i, which is the income rank of the ith individual in the population.

The income of any individual i, may now be expressed as a function of his rank, X_i,

$$Y_i=f(X_i) \tag{1}$$

It is assumed for simplicity that the function (f) is continuous.

By definition total income of the population is

$$Y=\int_0^N f(X)\,dX \tag{2}$$

where N is the size of the population.

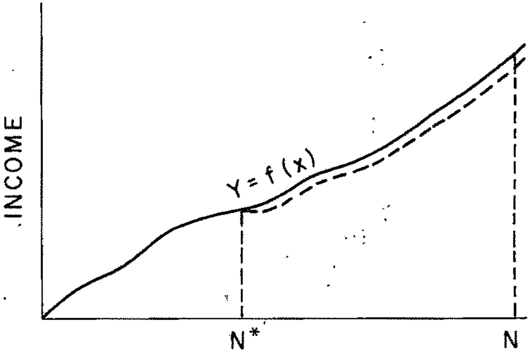
Diagramatically (see Figure 1) the income function is upward sloping, with income of the ith individual being the height of the curve at X_i, and total income being the area under the curve from the origin to N.

The income share of the bottom kth percentile is

$$s^{-k}=\frac{\int_0^{kN} f(X)\,dX}{\int_0^N f(X)\,dX} \tag{3}$$

Average income of the bottom kth percentile is

$$\bar{y}^{-k}=\frac{\int_0^{kN} f(X)\,dx}{kN} \tag{4}$$



RANK IN POPULATION

FIGURE 1. THE INCOME FUNCTION

Suppose that the population increases due to a new labor force entrant. Let Y^* be the income of the new entrant. Then $N^*=f^{-1}(Y^*)$ is the index number of the member of the initial population earning the same income as the new entrant.

Suppose that the income of all members of the original population is held constant. We now calculate the income shares and absolute incomes of the bottom kth percentile of the augmented population to see how both are affected by the new entrant. Keep in mind that no one's income has changed. Looking back at Figure 1, total income after the arrival of the new entrant is the area under the original curve plus Y^*.

Assigning the subscript zero for the base period and t for the period after the new entrant, we have the following t period shares for the bottom k percent of the population:

$$s_t^{-k}=\frac{\int_0^{k(N+1)} f(X)\,dX}{\int_0^N f(X)\,dX+f(N^*)} \tag{5}$$

for $N^*>k(N+1)$

$$s_t^{-k}=\frac{\int_0^{k(N+1)-1} f(X)\,dX+f(N^*)}{\int_0^N f(X)\,dX+f(N^*)} \tag{6}$$

for $N^*<k(N+1)$

[2] Note that I am taking a narrow view of the problem. In a more general context, population growth may also affect family size, education, and lifetime income profiles, all of which will also affect distribution statistics.

As the reader can see, if the new entrant falls outside the group whose share is being calculated, (equation (5)), then total income of the group rises, because population growth forces us to change the limits of integration. This may either increase or decrease the income share of the group depending on the relative rates of growth of overall average income and average income at the bottom. That in turn depends on the income level of the new entrant, that is, on N^*. It is easy to show the following facts regarding the relationship between s_t^{-k} and N^*.

$$\text{for} \quad N^* > k(N+1), \frac{\partial s_t^{-k}}{\partial N^*} < 0$$

$$\text{for} \quad N^* < k(N+1), \frac{\partial s_t^{-k}}{\partial N^*} > 0$$

$$\text{for} \quad N^* = 0, s_t^{-k} < s_0^{-k}$$

$$\text{for} \quad N^* = k(N+1), s_t^{-k} > s_0^{-k}$$

Consider now average income of the bottom kth percentile. It too is affected by population growth and by N^*. This can be seen in equations (7) and (8).

$$(7) \quad \bar{y}_t^{-k} = \frac{\int_0^{k(N+1)} f(X)\,dX}{k(N+1)}$$

for $N^* > k(N+1)$

$$(8) \quad \bar{y}_t^{-k} = \frac{\int_0^{k(N+1)-1} f(X)\,dX + f(N^*)}{k(N+1)}$$

for $N^* < k(N+1)$

From equations (7) and (8) it is clear that

$$\frac{\partial \bar{y}_t^{-k}}{\partial N^*} > 0, 0 < N^* \leqslant k(N+1)$$
$$= 0, N^* > k(N+1)$$

Furthermore:[3]

$$\bar{y}_t^{-k} < \bar{y}_0^{-k}, N^* < k(N+1) - 1$$
$$> \bar{y}_0^{-k}, N^* > kN$$

The expansion of the population causes a fall in apparent average income of the bottom kth percentile if the new entrants earnings fall inside the class limit, otherwise apparent average income increases. As with the income share, these changes are due to the necessary changes in the limits of integration due to population growth, not changes in real income of members of the bottom kth percentile.

It should be clear at this point that statistics like the share of income accruing to the poor, or the real income growth rate of the poor, are sensitive to growth in the income-earning population. Because of this there is an important distinction to be made between the poor, meaning the bottom of the distribution, and the base-period poor, by which I mean the people who were poor in some base year. Measured average income of the poor rises or falls as the population increases, even though the income of the base-period poor is held constant. If we are interested in knowing how the poor of the base period fared over time, and surely this is what the whole debate over trickle-down is all about, the real income growth rates of the bottom kth percentile will not tell us.

The points I have been making here about the poor apply with equal force to the rich. Reported average income of the rich will fall even though the income of the base-period rich stays constant, provided that the average wage of new entrants is less than the minimum income of the rich.[4] The behavior of the reported share of the rich is slightly more complicated. Clearly the share is a declining

[3] $$y_t^{-k} = \frac{\int_0^{k(N+1)-1} f(X)\,dX + (N^*)}{k(N+1)-1} < \frac{\int_0^{kN} f(X)\,dX}{kN}$$

if $f(k(N+1)-1) > \bar{y}_t^{-k}$, a condition almost sure to be satisfied.

[4] $N^* < (1-l)(N+1)$ where l is the top l percentile of the population.

TABLE 1—BRAZIL: COMPARISON OF INCOME DISTRIBUTION BY INCOME DECILES, 1960-70
(Income Earning Population)

Population Shares	Percentage of Income			Average Income (in 1970CR$ per Month)		
	1960	1970	Percent Change	1960	1970	Percent Change
Bottom 10 Percent	1.17	1.11	−5.1	25.	32.	+28.0
Bottom 40 Percent	11.57	10.00	−13.6	60.	71.	+18.3
Middle 40 Percent	34.08	27.80	−18.5	174.	197.	+11.9
Top 20 Percent	54.35	62.20	+14.5	560.	886.	+58.2
Top 10 Percent	39.66	47.79	+20.5	815.	1,360.	+66.9
Top 5 Percent	27.69	34.86	+25.9	1,131.	1,984.	+75.4
Top 1 Percent	12.11	14.57	+20.3	2,389.	4,147.	+73.6
Total	100.00	100.00		206.	282.	+36.9

Source: Langoni, p. 64.

function of N^*. Whether or not the t period share of the rich, s_t^{+l}, is greater or less than s_0^{+l} is indeterminate. If $N^*=0$, then the share of the rich will rise. If N^* falls close to the mean income of the base-period population, the share of the rich must fall.

II

Having made the general point that statistics such as the share or the growth rate of income are sensitive to population growth, we now want to see how quantitatively important the effect of population growth is by calculations for a particular country. I have chosen Brazil for this purpose because its distribution statistics are good, its case is well known, and it has a very rapid labor force growth rate.

For the 1960's, rising income inequality in Brazil has been widely noted and criticized. The data upon which that conclusion is based are shown in Table 1. As can be seen, average income growth rates at the bottom of the income pyramid lagged behind the average for the population as a whole. As a result the income share going to the poorest 10 or 40 percent fell between 1960 and 1970. At the top of the distribution, shares and relative incomes have the fastest rates of growth.

Section I showed what would have happened to shares and average incomes when a population's income is held constant as new entrants are inserted into the income pyramid. But in a real economy, things are not so simple. Income gains are losses accrued to the original population at the same time that the labor force is augmented by many new entrants with different incomes. In order to track the original labor force we have to be able to separate the new entrants from the survivors.

Fortunately we can make this separation for Brazil using age-specific income distributions for 1970. These allow us to decompose the 1970 labor force into survivors from 1960, and the new entrants over the decade, and to calculate a separate income distribution for each group. Details of the calculations are shown in the Appendix, and the resulting profile of the 1970 labor force is shown in Table 2. The income classes are those reported for 1970, and the income limit of each class was estimated by interpolation.

One can see from Table 2 that there were a lot of new entrants and that they were concentrated at the bottom of the 1970 income pyramid. Thirty-six percent of the 1970 labor force is made up of new entrants. Yet out of the 2.6 million Brazilians in the bottom decile, 1.4 million or 54 percent were new entrants. This same domination by new entrants is seen throughout all the bottom deciles of the 1970 distribution. This means that the distinction between the poor and the base-period poor will be an important one for Brazil. Anyone who compares incomes over time for the poor in Brazil is unwittingly comparing the incomes of two different groups with a high proportion of new

TABLE 2—SURVIVORS AND NEW ENTRANTS IN THE 1970 LABOR FORCE

	Labor Force Earning Less Than (1)	New Entrants Earning Less Than (2)	Survivors (col. (2)−col. (1))
<45CR$	2,607,975	1,398,616	1,209,359
<70	5,215,950	2,658,000	2;557,950
<96	7,823,975	3,851,000	3,972,925
<124	10,431,900	4,763,000	5,668,900
<157	13,039,875	5,694,000	7,345,875
<188	15,647,850	6,666,000	8,981,850
<241	18,255,825	7,592,000	10,663,825
<338	20,863,800	8,344,000	12,519,800
<720	23,471,775	9,097,688	14,374,087
<1160	24,775,763	9,278,688	15,499,075
>1160	1,303,987	113,478	1,188,509
Total	26,079,750	9,392,166	16,687,584

Note: The income distributions by age cohort shown in Langoni were graphed by linear approximation. New entrants defined as all of age agroups 10-19, plus the difference between the 1970 and 1960 labor force in age groups 20-24; 25-29; 30-39.

entrants. He most assuredly is not tracking the base-period poor.

Let us now calculate the growth rate of income of the base-period poor. This group will be the survivors of 1960's bottom 10 or 40 percent. If we think that income is a good indicator of welfare for the poor, this is the income growth rate we should be interested in. In order to make this calculation, we first must separate out the members of the 1960 labor force who retired between 1960 and 1970. Knowing that, we know how big a group, in 1970, the 1960 bottom decile or 40 percent was.

Since we have already calculated a distribution of new entrants, we know by subtraction the distribution of 1960's survivors. If we make the simplifying assumption of no downward mobility, it is then easy to estimate the 1970 income of any base-period group.

Consider the bottom decile. In 1970 it contained 2.6 million people. Subtracting the new entrants, we have an estimate of the group which was in the 1960 labor force (1.2 million). The 1960 bottom decile, however, had 1.9 million people; 200 thousand of those retired, leaving 1.7 million. This is the group of surviving poor whose real income gains we are trying to calculate. Now only 1.2 million of them were still in the bottom decile in 1970, which means that approximately 500,000 received salary increases large enough to push them above the 1970 income limit of the bottom decile (45CR$). Using the underlying distribution and assuming no downward mobility, we calculate the average income of this group at 49.5CR$.[5] With this information it is a simple matter to calculate total income of 1960's surviving poor. It is the income of those who remained in the bottom decile plus the income accruing to the 500 thousand above the upper-income limit. For simplicity, I relegate details to the Appendix, and show in Table 3 the resulting average incomes and growth rates for the bottom 10 and 40 percent. I did not make an equivalent calculation for the rich because of the difficulties in interpolation at the top of the distribution.

What is noteworthy about Table 3 is the wide divergence between the real income growth reported for the bottom 10 or 40 percent and the real gains enjoyed by the base-period poor. Rather than a 28 percent gain in real income, 1960's bottom decile had a 57 percent gain, a significant difference

[5] The assumption of no downward mobility means that the absolute income gains shown in Table 6 are, if anything, underestimates.

TABLE 3—REAL AND APPARENT INCOME OF THE POOR, 1960-70

	Per Capita Average Real Income (CR$)		Growth in Income 1960-70
	1960	1970	(Percent)
As Reported (Table 1)			
Bottom 10 Percent	25	32	28
Bottom 40 Percent	60	71	18
Overall	206	282	37
Survivors (Base-Period Poor)			
Bottom 10 Percent	25	39.2	57
Bottom 40 Percent	60	83.8	40
Overall	206	361	75

Note: Figures are monthly per capita income expressed in cruzeiors of 1970. See Appendix for calculation of survivors income.

TABLE 4—THE DISTRIBUTION OF INCOME OF SURVIVORS AND NEW ENTRANTS, 1970

	1960	1970
Gini[a]	.50	.56
Survivors' Gini	.50	.58
New Entrants' Gini	–	.516
Mean Income		
Survivors	206	361
New Entrants	–	193

[a] Differs slightly from Gini reported in Langoni because he had access to raw data. For comparability only my Ginis are shown.

both for their welfare and for any interpretation of the period. It is clear that for Brazil, one cannot ignore the measurement problem we have been investigating. Decile comparisons over time of the sort that have been done by most investigators will seriously underestimate the growth in income of the base-period poor in any economy with a rapidly growing labor force whose new entrants find their first jobs at the bottom of the income pyramid.

The reader might be wondering whether these estimates reverse the findings of virtually every observer that, over the decade, the rich gained relative to the poor. After all, my estimates of real income growth for the base-period poor are greater than the overall growth in average real income reported by Langoni as 37 percent. The answer, seen in Table 3, is no. The new entrant's correction raises the real income growth of all survivors by more than it does for the growth for the base-period poor. Thus, in Brazil, the poor did lose ground relative to the rich, just as the original distribution measures indicated. However, the correction shows that this took place along with an improvement in absolute income which is substantially higher than previously reported. The base-period poor may have lost ground, but their real income increased by 4.7 percent per year, not a bad record.

While the base-period poor apparently enjoyed substantial income growth, there must have been a simultaneous widening in income inequality among survivors. This can be shown by calculating survivor's and new entrants' Gini coefficients separately from the data underlying Table 2. (See Table 4.) As can be seen, the overall Gini *understates* the rise in inequality among the base-period population. From the statistical procedure involved in calculating the survivor's Gini, one would have expected the opposite. For by definition the survivors group eliminates teenagers, the bottom tail of the income distribution.[6] Hence income variance among survivors should diminish. That it did not is testimony to the high degree of inequality that accompanied income growth in Brazil.

[6] I am indebted to Alan Blinder for pointing out this bias in the survivors' Gini to me.

TABLE 5—PER CAPITA REAL INCOME AND GROWTH RATES BY AGE CLASS 1960-70

Age Class	1960CR$	1970CR$	Income Growth Rates Over Decade by Age (Percent)	Base Period Age Group (Percent)
10-14	52	58	11.5	280
15-19	101	109	7.9	176
20-29	134	235	22.7	85
30-39	243	341	40.3	58
40-49	251	385	53.4	41
40-59	249	355	42.6	25
60-69	217	312	43.8	–
70+	173	228	31.8	–

Source: Langoni, p. 86. Figures are average monthly income in cruzeiros of 1970.

III

Over the 1960's there was a very substantial widening in the wage differential across age groups. Entry level salaries grew slowly relative to those of the 30–49-year-olds as can be seen in Table 5. At the same time the composition of the labor force shifted towards a greater proportion of young workers. Both of these factors should increase measured inequality. Hence, a natural question is: how important were these two sources of inequality to the observed increase in the Gini coefficient? This is the question to be examined in this section.

Graham Pyatt in 1976 showed formally how the overall Gini could be split into three additive parts.[7] The first, G_A, measures the contribution of pure intracohort inequality; the second, G_B, inequality coming from overlaps between cohorts—that is, income differences between the richest teenagers and the poorest 40-year-olds; the third, G_C, the inequality coming from differences between the mean income of different cohorts. G_C, which Paglin called the age Gini, shows what inequality would be if each member of each cohort had the cohort mean income.

From this decomposition it might appear that G_C measures the contribution of intracohort inequality to overall inequality. However, that is not correct. Sheldon Danziger et al. and Keith Horner commenting respectively on Paglin and Pyatt point out that in fact the three sources of inequality are not independent. A rise in G_C because of a widening in age-wage differentials must reduce the overlap Gini G_B, if the distribution within each cohort (G_A) is held constant. For example, with a wider difference between the mean income of teenagers and the 40–49 age group, we should expect to find fewer teenagers with incomes higher than the poorest 40–49-year-olds. Hence $G_A + G_B$, the Paglin Gini, is sensitive to the mean income of each cohort. It is also sensitive to the distribution of households by cohort.

To avoid the interdependence problems, I proceed in a different fashion.[8] Instead of looking directly at G_A, G_B, and G_C let us construct a series of hypothetical income distributions, each of which holds constant all but one of the possible sources of inequality. (See Table 6.) Columns (1) and (2) show the actual distributions. Column (3) shows what the distribution would have been in 1970 with the 1970 intracohort distribution but with the 1960 distribution of individuals by cohort and age-income profile. A comparison of column (3) and (1) shows the effect of changes in intracohort inequality.

[7] Paglin (1975), working independently, proposed a quite similar decomposition which he used to investigate the contribution of a change in the age-income profile to *U.S.* inequality.

[8] My procedure follows closely that of Danziger et al. except that, having only the 1970 age-specific income distributions, I was unable to show the hypothetical 1970 distribution with 1960 cohort distributions.

TABLE 6—PYATT-GINIS FOR 1960-70

	Actual		1970 Gini Holding Constant at the 1960 Level:	
			Mean Income by Cohort and Population Weights	Population Weights
	1960 (1)	1970 (2)	(3)	(4)
G_A Intracohort	.0796	.0887	.0898	.0910
G_B Overlap	.2871	.2919	.3243	.2940
G_C Age	.1306	.1856	.1306	.1728
Overall Gini	.4973	.5662	.5447	.5578

Source: Computations by Author from age-specific income distributions for 1970 in Langoni.
Note: Col. (3): 1960 cohort mean income, 1970 population weights by cohort, 1970 cohort income distribution.

Column (4) shows the hypothetical 1970 distribution with the 1970 intracohort distributions and 1970 age-income profile. A comparison of columns (4) and (3) measures the influence of changes in the age income profile.

Finally column (2), the actual 1970 distribution, allows the population weights to take their 1970 values. Comparing columns (4) and (2) shows the effect of changes in the population structure on overall inequality. Note that this procedure gives only an approximate estimate of the separate effect of the three sources of inequality because of the possible interaction between effects in switching between one population base and another.

Let us now attempt to interpret the table. Looking first at columns (1) and (2), note that 80 percent of the increase in the overall Gini comes from the 5.5 percentage point increase in the age Gini. This is a result of the significant widening in the age-wage profile. A superficial interpretation of the data would conclude that most of the rising inequality in Brazil came from this source. But that is wrong. Compare columns (3) and (4). Here everything is held constant but the age-income profile which takes its 1960 value in column (3) and its 1970 value in column (4). As can be seen, the change does increase inequality, but the effect is small. The reason this measurement has such a different implication from the simple comparison of the age Gini's in columns (1) and (2) is because of the dependence of G_B, the overlap term, on the age-income profile. Changing that profile leads to a big rise in G_C (from .1306 to .1728), matched by an almost equally large decline in G_B (from .3243 to .2940), leaving the overall Gini relatively unchanged.

Clearly the main contributor to rising inequality is the change in the intracohort distributions (compare cols. (1) and (3)). This source alone raises the overall Gini by almost 5 points. The reason this does not show up in the simple comparison of $G_A + G_B$ proposed by Paglin is again because of the overlap term, G_B. The rise in intracohort inequality raises G_B by just about the same amount that the rise in intercohort inequality reduces it, with the result that G_B stays almost constant, appearing to contribute little to the large rise in inequality.

Population growth also adds to inequality. (Compare cols. (4) and (2).) However, it is the least important of the three sources, adding less than one percentage point to the overall Gini.

Once again we find evidence of the inequality of growth "Brazilian style." Brazil's was a growth model which raised incomes at the bottom substantially but only at the cost of a significant rise in inequality. While inequality increased, the data do not support the conclusion that a particular group remain on the bottom of the income pyramid their entire lives. The bottom of the income pyramid is composed primarily of teenagers. Column (5) of Table 5 shows clearly that the income of the average teenager increased rapidly over the 1960's. Just because the average salary of teenagers did not grow rapidly does not mean that the income of a particular group of teenagers did not. The opposite is true for Brazil. Thus significant upward mobility and rising inequality can and did occur together in Brazil.

IV. Conclusion

I have shown that population growth has a substantial effect on measured income growth and income share for both the rich and the poor. It forces us to distinguish between the base-period rich or poor, and the reported rich or poor. For Brazil, actual income data were used to show that the population effect is quantitatively important. New entrants partially mask two conflicting features of recent Brazilian growth. Absolute income growth for the base-period poor or base-period teenager was larger than reported for the poor or for teenagers, but so also was the rise in inequality for the base-period population.

It is beyond the scope of this paper to consider in detail the reasons for rising inequality. Nonetheless I was able to show that almost all the rise in the Gini coefficient results from the rise in intracohort inequality rather than changes in the age-income profile as might be supposed from a superficial decomposition of the Gini. The general conclusion is that growth in Brazil was even more regressive than had been thought, but it was not as immiserating as many have claimed. Significant upward mobility *and* rising inequality can and did occur together. The distribution measures currently in vogue, ignoring the effect of changing population, reflect the second, but not the first.

APPENDIX

Bottom 10 Percent (earning<45CR$)	$\bar{y}$	Number	Total y
(1) As reported	32.69	2,607,974	85,254,670
New entrants total:			
10-14	27.47	308,799	8,481,668
15-19	30.48	619,817	18,891,366
20-24	32.91	340,000	11,189,745
25-29	35.86	110,000	3,944.785
30-39	36.00	20,000	720,000
(2) Total	30.91	1,398,616	43,227,564
(3) Survivors (1)-(2)	34.75	1.209,359	42,027,106
(4) 1960 Workers in interval 45-54CR$	49.5	524,337	25,954,682
(5) 1960 Survivors (3)+(4)	39.2	1,733,696	67,981,788
Bottom 40 Percent (earning <124CR$)			
(1) As reported	70.7	10,431,897	737,262,729
New Entrants			
10-14	44.7	543,000	24,278,768
15-19	64.6	2,080,000	134,343,582
20-24	71.9	1,495,000	107,415,750
25-29	73.8	510,000	37,620,899
30-39	76.3	114,000	8,698,200
(2) Total	65.9	4,742,000	312,357,199
(3) Survivors (1)-(2)	74.7	5,698,897	424,905,530
(4) 1960 Workers in interval 124-146CR$	135.	1,018,071	137,440,000
(5) 1960 Survivors (3)+(4)	83.8	6,707,968	562,345,530

Retirement Calculations:

	Age classes				
	40-49	50-59	60-69	+70	Total
Bottom 10 Percent (<45CR$)	30,000	62,500	61,001	44,902	198,404
Bottom 40 Percent (<124CR$)	237,500	310,000	280,612	192,440	1,020,552

To calculate retirements I assumed that the 1970 and 1960 distribution of retirees was equal. I thus applied the proportion of the 1970 age group earning less than 45CR$ to the numbers of this age group in 1960 who retired over the decade.

REFERENCES

E. Bacha and L. Taylor, "Brazilian Income Distribution in the 1960's: 'Facts,' Model Results and the Controversy." *J. Develop. Stud.*, Apr. 1978, *14*, 271–97.

S. Danziger, R. Haveman, and E. Smolensky, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 505–12.

G. Fields, "Who Benefits from Economic Development?—A Reexamination of Brazilian Growth in the 1960's," *Amer. Econ. Rev.*, Sept. 1977, *67*, 570–82.

A. Fishlow, "Brazilian Size Distribution of Income," *Amer. Econ. Rev.*, May 1972, *62*, 391–4102.

______, "Brazilian Income Distribution: Does Trickle-Down Really Work?," unpublished paper, World Bank 1977.

K. Horner, "Interpreting Pyatt's Decomposition of the Gini Coefficient: A Comment," mimeo., Nat. Health and Welfare, Canada, undated.

S. Kuznets, "Demographic Aspects of the size Distribution of Income: An Exploratory Essay," *Econ. Develop. Cult. Change*, Oct. 1976, *25*, 1–94.

Carlos G. Langoni, *Distribuica da Renda e Desenvolvimento Economico do Brasil*, Rio de Janeiro 1973.

S. A. Morley, "Growth and Inequality in Brazil," *Luso-Brazilian Rev.*, Winter 1978, *15*, 244–71.

______ **and J. G. Williamson,** "Growth, Wage Policy and Inequality: Brazil During the Sixties," workshop paper no. 7519, SSRI, Univ. Wisconsin, July 1975.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

______, "The Measurement and Trend of Inequality: Reply," *Amer. Econ. Rev.*, June 1977, *67*, 520–31.

G. Pyatt, "On the Interpretation and Disaggregation of Gini Coefficients," *Econ. J.*, June 1976, *86*, 243–55.

G. S. Sahota, "Theories of Personal Income Distribution: A Survey," *J. Econ. Lit.*, Mar. 1978, *16*, 1–55.

Ricardo Tolipan and Arthur C. Tinelli, *A Controversia sobre Distribicao de Renda e Desenvolvimento*, Rio de Janeiro 1975.

J. Wells, "Distribution of Earnings, Growth and the Structure of Demand in Brazil during the 1960's" *World Develop.*, Jan. 1974, *2*, 9–24.

Are Market Forecasts Rational?

By FREDERIC S. MISHKIN*

This paper conducts tests of the rationality of both inflation and short-term interest rate forecasts in the bond market. These tests are developed with the theory of efficient markets and make use of security price data to infer information on market expectations. A closer look at whether market forecasts of inflation and interest rates are rational seems necessary because of recent work (see James Pesando, John Carlson, Donald Mullineaux, and Benjamin Friedman) which has evaluated the inflation and interest rate forecasts from the Joseph Livingston and Goldsmith-Nagan surveys. A common empirical result in these studies is that the survey forecasts are inconsistent with the restrictions implied by the theory of rational expectations. What conclusions about the behavior of market expectations should we draw from these results?

One view which associates survey forecasts with those of market forecasts would take these empirical results as evidence that the market is not exploiting all information in generating its forecasts. Friedman's results are particularly disturbing in regard to the possible irrationality of the bond market because this study uses data from the Goldsmith-Nagan interest rate survey which is made up of interest rate forecasts from *actual* participants in that market.

An alternative view would hold that markets probably do display rationality of expectations. Irrationality in the Livingston and Goldsmith-Nagan survey data would then indicate that these data cannot be used in empirical work to describe market expectations.[1]

There are two reasons why the latter view receives support. Survey data are frequently believed to be inaccurate reflections of market participants' behavior and are thus considered to be unreliable. Of even greater importance is a point that is often ignored in discussing the properties of expectations. *Not all market participants have to be rational in order for a market to display rational expectations.* The behavior of a market is not necessarily the same as the behavior of the average individual. As long as unexploited profit opportunities are eliminated by some participants in a market (this is analogous to an arbitrage condition), then the market will behave as though expectations are rational despite irrational participants in that market.[2] Therefore, survey forecasts do not necessarily describe the forecasts inherent in market behavior, and irrationality of survey forecasts does not in itself imply that market forecasts are also irrational.

One purpose of this paper is to provide indirect evidence on the usefulness of survey data like Livingston's and Goldsmith-Nagan's for describing the expectations reflected by markets. In particular, this paper conducts more direct tests of the rationality of the bond market's interest rate and inflation forecasts, and these tests are similar to those conducted in the studies mentioned in the opening paragraph. Because these tests are designed to use actual price data to infer information on market expectations rather than relying on survey data, they can provide direct information on the rationality of a particular market, such as the bond market. This permits a clearer interpretation of results which indicate irrationality in survey forecasts. The empirical work in this paper will thus shed light not only on the usefulness of these surveys for other empirical research, but also on the rationality of expectations in markets such as those in which bonds are traded.

*Department of economics, University of Chicago. I would like to thank Andrew Abel, Eugene Fama, David Galenson, Robert Lucas, Merton Miller, Thomas Sargent, and the participants in the Money Workshop, and the Econometrics and Statistics Colloquium at the University of Chicago for their helpful comments.

[1] See Pesando, for example.

[2] See my 1978 paper and the following section for a discussion of this issue.

I. Tests of Forecast Rationality

The theory of rational expectations, initially developed by John Muth, asserts that both firms and individuals, as rational agents, have expectations that are optimal forecasts using all available information. Rationality of expectations requires that

$$E(X_t - X_t^e|\phi_{t-1})=0 \tag{1}$$

where X_t^e is the one-period-ahead forecast of a variable X_t, generated at the end of period $t-1$, and ϕ_{t-1} is the set of information available at the end of $t-1$. This implies that the forecast error $X_t - X_t^e$ should be uncorrelated with any information or linear combinations of information in ϕ_{t-1}.

This implication is the basis of the tests of rationality found in the studies of survey forecasts mentioned above. Consider the following equations:

$$X_t = b_0 + \sum_{i=1}^{k} b_i X_{t-i} + u_{lt} \tag{2}$$

$$X_t^e = c_0 + \sum_{i=1}^{k} c_i X_{t-i} + u_{2t} \tag{3}$$

These equations can be estimated with ordinary least squares (*OLS*), under the assumption that $E(u_{lt}|\phi_{t-1})=E(u_{2t}|\phi_{t-1})=0$ (implying that the u's are serially uncorrelated and uncorrelated with the X_{t-i}). Under the hypothesis of rational expectations, the estimated b_i coefficients should not differ from the estimated c_i coefficients except by chance.[3] This null hypothesis that

$$b_i = c_i \text{ for all } i=0,\ldots,k \tag{4}$$

is tested in the studies of survey forecasts with a conventional F-test.

The rationale behind this test becomes more obvious by subtracting (3) from (2) to obtain

$$X_t - X_t^e = (b_0 - c_0) + \sum_{i=1}^{k}(b_i - c_i)X_{t-i} + (u_{1t} - u_{2t}) \tag{5}$$

The rationality criterion in (1) combined with $E(u_{lt}|\phi_{t-1})=E(u_{2t}|\phi_{t-1})=0$ implies the null hypothesis $b_i = c_i$ for all i. Since the *OLS* estimates of $b_i - c_i$ in (5) are numerically equal to the *OLS* estimates of b_i in (2) minus the *OLS* estimates of c_i in (3), these separate estimates of b_i and c_i should be equal, except for statistical variation. Note that even if other information besides the k lagged values of X is used to forecast X, it is clear from (5) and (1) that the test of these cross-equation rationality restrictions is still valid.[4] However, because $E(u_{lt}|\phi_{t-1})$ and $E(u_{2t}|\phi_{t-1})$ need not equal zero in this case, the u's could be correlated with lagged X's. Then the estimated b_i and c_i coefficients would not be consistent; yet the rationality restrictions would hold because these estimated coefficients would suffer from identical bias.[5]

The theory of efficient markets leads to restrictions that are similar to those in (4) which can also be easily tested. Market efficiency implies that securities' prices in a capital market should reflect all available information, and hence an expectation assessed by the market should equal the true expectation conditioned on all available information, $E(\ldots|\phi_{t-1})$.[6] In order to give this concept empirical content, we must specify the relationship between the probability distribution of future prices and current prices.

[3] Franco Modigliani and Robert Shiller pointed out the rationality principle tested here: that one-period-ahead forecasts of a variable and the realizations of that variable should have the same regression relationship to past realizations of that variable.

[4] The test is valid in the sense that, except for chance, a rejection of the null hypothesis can occur only if expectations are not rational. However, a failure to reject the null hypothesis, even asymptotically, does not rule out irrationality. See the paper by Andrew Abel and myself for a more extensive discussion of this point.

[5] Another way of stating this point is to say that rational expectations by itself has no implications about the right-hand side variables used in (2) or (3) or about the properties of the u_{1t} and u_{2t} error terms. It only has implications about the equality of the *OLS* estimated coefficients in (2) and (3).

[6] See Eugene Fama (1976a) for a more detailed treatment of the theory of efficient markets.

This requires a model which describes how current equilibrium prices are determined. Here, the market is assumed to equate expected, one-period, holding returns across securities, allowing for liquidity (risk) premiums which are constant over time.

For example, in the case of long-term bonds, the one-period return $BRET_t$ is the nominal return from holding the long-term bond from $t-1$ to t which includes both capital gains plus interest payments. The model of market equilibrium implies that

$$(6) \qquad E_m(BRET_t|\phi_{t-1})=r_{t-1}+\delta$$

where r_{t-1} = the return on a one-period bond (which of course equals the expected one-period return)—this is just the short-term interest rate, δ = the constant liquidity (risk) premium, and $E_m(\ldots|\phi_{t-1})$ = expectation assessed by the market at $t-1$. Market efficiency then implies that

$$(7) \quad E(BRET_t-E_m(BRET_t|\phi_{t-1})|\phi_{t-1}) = E(BRET_t-r_{t-1}-\delta|\phi_{t-1})=0$$

If we denote the equilibrium return of $r_{t-1}+\delta$ as a "normal" return, then the equation above states that no unexploited profit opportunities exist in the bond market: at today's price, market participants cannot expect to earn a higher than normal return by investing in a long-term bond. The efficient markets equation (7) is analogous to an arbitrage condition. Arbitrageurs who are willing to speculate may perceive unexploited profit opportunities and purchase or sell bonds until the price is driven to the point where (7) holds.[7] Thus market efficiency does not require that *all* participants in the market are rational and use information efficiently.

Equation (7) above implies that only when new information hits the market will $BRET_t$ differ from $r_{t-1}+\delta$. This is equivalent to the proposition that only unanticipated movements (surprises) in variables can be correlated with $BRET_t-r_{t-1}$. This leads to the following efficient markets model:

$$(8) \quad BRET_t-r_{t-1}=\delta+(X_t-X_t^e)\alpha+\varepsilon_t$$

where an e superscript denotes expected values conditional on all past available information (i.e., $X_t^e=E_m(X_t|\phi_{t-1})$, a one-period-ahead optimal forecast), X_t = a variable (or vector of variables) relevant to the pricing of long bonds, α = a coefficient (or vector of coefficients) and ε_t = an error process where $E(\varepsilon_t|\phi_{t-1})=0$ and hence ε_t is serially uncorrelated.[8] The distinction in equation (8) between the possible effects from unanticipated vs. anticipated movements in variables is indeed an important feature of recent empirical work (see, for example, Robert Barro, 1977, 1978).

The assumption that the coefficient on r_{t-1} equals one in equation (6) has been subjected to empirical test in work by Fama and G. William Schwert and my 1978 paper and is not rejected.[9] Furthermore, as is discussed in

[7] For reasons discussed in Sanford Grossman and Joseph Stiglitz, the arbitrage-type condition of (7) cannot hold exactly because there must be compensation for risk taking, information collection and transactions costs. However, this in no way denies its usefulness as an approximation in the case where these costs are small relative to the size of market transactions, as we would expect to be the case for the interest rate and inflation data analyzed here.

[8] It is easy to show that this efficient markets model is consistent with the expectations hypothesis of the term structure where predictions of future short-term interest rates are optimal forecasts. To be more concrete, if the long-term bond is a discount security where the liquidity premium is a constant δ, the long interest rate RL_t is approximated by $RL_t=\delta+(1/n)[r_t^e+r_{t+1}^e+\ldots+r_{t+n-1}^e]$. When expectations of future short rates in this equation are optimally formed, or equivalently are "rational" in the sense of Muth, then the expectations hypothesis described by the equation above leads to the same implications as equation (8) in the text. Note also that the efficient markets model does not imply causation from $X_t-X_t^e$ to $BRET_t-r_{t-1}$. It is equally plausible that causation runs in the other direction or that a third factor affects both of these variables simultaneously.

[9] This assumption was also tested using the 1954–76 sample period. A quarterly bond returns series was regressed on the beginning of period, ninety-day Treasury bill rate (also at quarterly rates) using weighted least squares to correct for heteroscedasticity. (My 1978 paper describes this procedure.) The coefficient on the bill rate was not significantly different from one at the 5 percent level (t=.51). In a recent paper, Shiller has found evidence which can be interpreted as implying

Fama (1976a), as long as $E_m(BRET_t|\phi_{t-1})$ has small variation relative to other sources of variation in the actual returns—and this appears to be the case for the long-term bonds discussed here[10]—assumptions describing the equilibrium return are not critical to empirical tests of the efficient markets model.[11]

Substituting expectations of X from equation (3) into (8) we have an efficient markets model of the following form:

$$(9)\quad BRET_t - r_{t-1} = \delta + \alpha\left(X_t - \left(c_0 + \sum_{i=1}^{k} c_i X_{t-i}\right)\right) + \varepsilon_t'$$

where $\varepsilon_t' = \varepsilon_t - \alpha u_{2t}$

Equations (9) and (2) can then be stacked into one regression system, and it can be estimated by non-linear least squares methods imposing the restrictions in (4) implied by forecast rationality: that $b_i = c_i$ for all i. In order to obtain more efficient parameter estimates as well as consistent test statistics, weighted least squares corrections must be made for heteroscedasticity both within and across equations in the system.[12] Furthermore, avoiding a time-aggregation problem when averaged data is used in tests of rationality requires a bond return series which is constructed from end-of-period data.[13] The rationality restrictions can now be tested in the efficient markets framework above with a simple likelihood ratio test. However, note, that here these restrictions have been generated under the maintained hypothesis that $BRET_t - r_{t-1}$ is correlated with only contemporaneous surprises.[14] The likelihood ratio statistic $-2\,log(L^c/L^u)$ is distributed asymptotically as $\chi^2(q)$ where q is the number of non-linear constraints (which equals k in the equations (9) and (2) system), and L^c = likelihood of the constrained system, L^u = likelihood of the unconstrained system. In this non-linear least squares system, the likelihood ratio statistic is just:[15]

$$2n(log(SSR^c) - log(SSR^u))$$

where

SSR^c = sum of squared residuals from the constrained system,

SSR^u = sum of squared residuals from the unconstrained system, and

that the liquidity premium is correlated with the spread between long rates and short rates. To test this proposition for the 1954–76 sample period, $BRET_t - r_{t-1}$ was regressed on this spread, again using weighted least squares to correct for heteroscedasticity. The evidence supporting Shiller's proposition is even weaker in this sample period than was true in the regression results reported in my 1978 paper: the coefficient on the spread variable was not significantly different from zero at even the 10 percent significance level ($t = 1.01$).

[10] For example, using the model of market equilibrium described in the text, over the 1954–76 period the variation in $E_m(BRET_t|\phi_{t-1})$ is less than 2 percent of the variation in the actual return stemming from other sources.

[11] A more precise wording of this point would state that in the case discussed here, tests of hypotheses concerning the equilibrium return would have very low statistical power. This is essentially the same point made by Charles Nelson and Schwert in their comment on Fama (1975). Note also that more discriminating tests provide evidence that liquidity premiums are not constant over time. See, for example, Fama (1976b).

[12] The following iterative procedure was used to correct for heteroscedasticity in these initial estimates. In the estimates of each equation, if Goldfeld-Quandt (1965) tests indicated that heteroscedasticity existed within an equation, then the variables in this equation were weighted using a time-trend procedure outlined in H. Glesjer. In addition, the variables in each equation in the system were appropriately weighted so that each equation individually had the same sum of squared residuals. The system was then estimated jointly, and the resulting sum of squared residuals for each equation were then used to weight the variables in each equation so that each equation would again have the same sum of squares. The non-linear system was then estimated again. The resulting sums of squared residuals were now so close to being equal that no further iterations were performed. Indeed, further iterations produced only very slight changes in parameter estimates and in the likelihood ratio statistics reported here.

[13] See Holbrook Working and the discussion of the Franco Modigliani and Shiller paper in the following section.

[14] Clearly, if this maintained hypothesis which arises from the efficient markets model were invalid, this test could lead to rejection of these restrictions even if rationality were valid. This issue is analyzed empirically in fn. 21.

[15] See Stephen Goldfeld and Richard Quandt (1972). Note that the same weights used for the heteroscedasticity corrections in the constrained system (see fn. 12) are used in the unconstrained system.

n = the number of sample period observations— thus $2n$ is the number of observations in the stacked regression.

As is shown in the paper by Abel and myself, this likelihood ratio test has the attractive property that it is a valid test of rationality under very general conditions.[16] Furthermore the rationality test proposed here is demonstrated to be asymptotically equivalent to a common test of market efficiency frequently used in the literature.[17]

III. Empirical Results

The first set of tests to be conducted here will scrutinize Friedman's result that the survey measures of interest rate forecasts are inconsistent with rationality. Friedman's results were obtained using thirty quarterly observations extending from September 1969 to December 1976, and this sample period is used to estimate the equations (9) and (2) system using bond return and Treasury bill rate data described in the Data Appendix. His choice of six lagged quarters in his autoregressive specification will also be used in these tests. An additional test will be conducted over the longer 1954–76 sample period to provide more information on the rationality of the bond market's forecasts.

Tests of the rationality of inflation forecasts will also be conducted in a similar manner using the non-linear efficient markets procedure. The 1959–69 sample period used by Pesando, Carlson, and Mullineaux, where many rejections of rationality have been found, will be used in these tests, as well as the longer 1954–76 sample period. Here, the Consumer Price Index (*CPI*) will be used to calculate the inflation rate and this data is also discussed in the Data Appendix.

A. *Results on the Rationality of Interest Rate Forecasts*

Table 1 provides the tests for the rationality of forecasts in the bond market using both Friedman's 1969–76 sample period and the longer 1954–76 sample period; while Table 2 provides the parameter estimates of the constrained efficient markets model using both sample periods.The p-values in Table 1 are the probability of obtaining that value of the likelihood ratio statistic or higher, under the null hypothesis that the rationality constraints are valid. A p-value less than .05 would indicate a rejection at the 5 percent level of the null hypothesis and, therefore, a rejection of forecast rationality in the bond market.

As the likelihood ratio statistics in Table 1 indicate, there is very little evidence in the bond market data supporting irrationality of interest rate forecasts. Not only are there no significant rejections of the rationality restrictions in either Friedman's sample period or the longer 1954–76 sample period, but the p-values of Table 1 are quite high. In addition, the efficient markets model from which these likelihood ratio statistics have been derived, whose parameter estimates are found in Table 2, has several attractive properties. The coefficients on the unanticipated movements of the bill rate are significantly different from zero at the 1 percent level, thus indicating that movements in short-term interest rates embody relevant information to the pricing of long-term bonds. As might be

TABLE 1—TEST OF FORECAST RATIONALITY: INTEREST RATES

	Sample Period	
	1969:3 to 1976:4	1954:1 to 1976:4
Likelihood Ratio Statistic	6.55	4.96
p-Value	.364	.549

Note: Likelihood ratio statistic is distributed asymptotically as $\chi^2(6)$. The p-value is the probability of finding that value of the likelihood ratio statistic or higher under the null hypothesis that the non-linear constraints are satisfied.

[16] The test is valid in the sense described in fn. 4. For example, if $u_{2t} \neq 0$ so that there would be errors in variables bias in the estimated α coefficient, the test is still valid. Correlations of $X_t - X_t^e$ with ε_t also leads to inconsistent estimates of α yet it again does not invalidate the likelihood ratio test for rationality.

[17] Yet, as we shall see, the tests conducted here do yield more information than the more common test.

TABLE 2—NON-LINEAR ESTIMATES OF THE EFFICIENT MARKETS MODEL

$$BRET_t - r_{t-1} = \delta + \alpha\Big(r_t - b_0 - \sum_{i=1}^{6} b_i r_{t-i}\Big) + \varepsilon'_t$$

$$r_t = b_0 + \sum_{i=1}^{6} b_i r_{t-i} + u_t$$

	Sample period	
	1969:3 to 1976:4	1954:1 to 1976:4
δ	.0055	−.0018
	(.0091)	(.0032)
α	−13.4452	−12.3800
	(4.6568)	(1.8264)
b_0	.0060	.0006
	(.0023)	(.0003)
b_1	.6158	1.0706
	(.1750)	(.0869)
b_2	.0639	−.3123
	(.1913)	(.1287)
b_3	.3159	.2189
	(.1869)	(.1331)
b_4	−.1434	.0296
	(.1872)	(.1348)
b_5	−.3195	−.1473
	(.1911)	(.1324)
b_6	.0463	.0906
	(.1790)	(.0909)

Note: Asymptotic standard errors are shown in parentheses; $BRET_t$ = quarterly bond return at quarterly rate; r_t = Treasury bill rate at a quarterly rate.

expected from the expectations hypothesis of the term structure, the sign of this coefficient is negative, indicating that an unanticipated rise in the bill rate is accompanied by higher long-term rates with a resulting lower bond return. Furthermore, the magnitude of this coefficient is quite close to that found in my 1978 paper.[18]

The failure to reject the rationality of interest rate forecasts in the bond market provides some resolution of how to interpret Friedman's result that the Goldsmith-Nagan survey measures of interest rate forecasts are irrational. The evidence here supports the view that the survey measures of interest rate forecasts are not an accurate description of the actual bond market forecasts. The use of these survey measures is thus suspect in other empirical work. The evidence also does not support the view that the bond market could have improved its forecasting behavior by more efficiently exploiting the information in the past bill rate movements. Of course, these results should not be surprising considering the large body of evidence which supports efficiency in the bond market.[19]

B. *Results on the Rationality of Inflation Forecasts*

The test of the rationality of inflation forecasts in the bond market can be found in Table 3, while the parameter estimates of the constrained efficient-markets model are in Table 4. The efficient-markets model does yield the expected result that an unanticipated rise in inflation is associated with higher long rates and lower bond returns—although the coefficients on unanticipated inflation are not as significant as were the coefficients on unanticipated interest rate movements. However, the likelihood ratio test rejects the rationality restrictions for the 1959–69 sample period at the 1 percent significance level,[20] and this is the sample period where other studies (see Pesando, Carlson, and Mullineaux) have also found the Livingston price expectations data to be irrational.[21]

[18] Note that in my 1978 paper I used Treasury bill data which is at an annual rate. Thus the coefficient on the unanticipated bill rate in that case must be multiplied by four when compared to the α coefficients in Table 1.

[19] See Fama's (1970) survey and the more recent work of Thomas Sargent and myself (1978, 1981).

[20] Because this rejection of rationality was so striking and therefore should be checked out, I performed a standard test of bond market efficiency, similar to those in my 1978 paper where I regressed $BRET_t - r_{t-1}$ on six lagged values of the inflation rate. The results for the 1959–69 sample period were similar to those of Table 3. The restrictions imposed by market efficiency (rationality) were rejected at the 1 percent significance level: $F(6,37) = 5.26$ while the critical F at 1 percent is 3.78.

[21] Because the rationality restrictions are generated under the maintained hypothesis that $BRET_t - r_{t-1}$ is uncorrelated with anticipated movements of X, the rejection here might arise from the invalidity of the maintained hypothesis and not from the irrationality of inflation expectations. To explore this possibility, the hypothesis of rational expectations can be tested using the techniques discussed in the text without maintaining the hypothesis that $BRET_t - r_{t-1}$ is uncorrelated with

TABLE 3—TEST OF FORECAST RATIONALITY: INFLATION

	Sample Period	
	1959:1 to1969:4	1954:1 to 1976:4
Likelihood Ratio Statistic	23.77	8.70
p-Value	.001	.191

Note: See Table 1.

TABLE 4—NON-LINEAR ESTIMATES OF THE EFFICIENT MARKETS MODEL

$$BRET_t - r_{t-1} = \delta + \alpha(\pi_t - b_0 - \sum_{i=1}^{6} b_i \pi_{t-i}) + \varepsilon'_t$$

$$\pi_t = b_0 + \sum_{i=1}^{6} b_i \pi_{t-i} + u_t$$

	Sample Period	
	1959:1 to 1969:4	1954:1 to 1976:4
δ	−.0036	−.0019
	(.0034)	(.0032)
α	−2.5189	−1.8685
	(1.3319)	(.8436)
b_0	.0003	.0012
	(.0008)	(.0006)
b_1	−.0464	.3778
	(.1461)	(.1031)
b_2	.6047	.5173
	(.1210)	(.1100)
b_3	.2626	.2075
	(.1497)	(.1224)
b_4	−.0477	−.1555
	(.1206)	(.1219)
b_5	.2104	−.0392
	(.1147)	(.1100)
b_6	.1233	−.0374
	(.1242)	(.1035)

Note: See Table 2. π_t = inflation rate at a quarterly rate measured by the change in the *log*(*CPI*) over the quarter.

A look at the unconstrained estimates of the autoregressive model of inflation and the efficient-markets model provides a clue as to why this rejection of rationality occurs. The sum of the coefficients on the lagged inflation rates in the autoregressive model of inflation is positive and greater than one, indicating that a rise in inflation would persist. On the other hand, the sum of these autoregressive parameters derived from the unconstrained efficient-markets model is negative, indicating that the bond market expected that a rise in inflation would be reversed.[22] This discrepancy is what leads to the rejection of the rationality of the bond market's forecasts of inflation, and it should not be all that surprising considering the sample period chosen. This sample period started with a low inflation rate which then rose to unusually high levels by the end of this period. The fact that this was an unusual period might then be the cause of the rejection of the rationality restrictions found in Table 3, even though the bond market would normally have rational inflation forecasts. A similar problem has been found for the rationality of inflation forecasts (represented by forecasts of exchange rate changes) in the German hyperinflation, again an unusual inflationary episode (see Jacob Frenkel). The likelihood ratio test on the rationality of the inflation forecasts in the longer 1954–76 period does provide some evidence supporting this conjecture. In this period there is no rejection of the rationality restrictions at the

X_t^e. This involves estimating the system

$$X_t = b_0 + \sum_{i=1}^{k} b_i X_{t-i} + u_{1t}$$

$$BRET_t - r_{t-1} = \delta + \alpha\left[X_t - \left(c_0 + \sum_{i=1}^{k} c_i X_{t-i}\right)\right]$$

$$+\theta\left[c_0 + \sum_{i=1}^{k} c_i X_{t-i}\right] + \varepsilon_t$$

and testing the null hypothesis that $b_i = c_i$ for all i. Note that this procedure involves a test of $k-1$ restrictions, one less than in the previous tests. When this test for rationality of the inflation forecasts was conducted using the same 1959–69 sample period, the rationality restrictions were still strongly rejected by the data. The resulting likelihood ratio statistic (distributed asymptotically as $\chi^2(5)$) equaled 16.65 with a p-value of .005. This rejection at the 1 percent level adds additional support to the view that inflation forecasts were not rational for this sample period.

[22] In the unconstrained autoregressive model of inflation, the coefficients ($\hat{b}_i$) of the lagged inflation rates are, starting with lag one: −.06, .59, .19, −.03, .30, and .25. In the unconstrained efficient markets model, the estimated $\hat{\beta}$ equals −1.52 and the implied coefficients ($\hat{c}_i$) of the lagged inflation rates for the expectations equation are, starting with lag one: −.27, .25, 1.04, −.30, −.94, and −1.60.

5 percent significance level. Thus it appears that the bond market may have had rational inflation forecasts when a longer time horizon is taken into account.[23]

What do these results tell us about the accuracy of the Livingston price expectations data? We must be somewhat careful in our interpretation of these results because the Livingston survey does not specifically sample those who are participants in the bond market, yet the following conclusion does seem to be indicated. Because the 1959–69 period is one in which the inflation forecasts in the bond market do not satisfy restrictions implied by rationality, the failure of survey measures to satisfy these restrictions cannot be taken as evidence that they are inaccurate measures of market expectations. Clearly, further research evaluating the rationality of the Livingston price expectations data using longer sample periods than the 1959–69 period is needed before we can pronounce on their accuracy.

C. Joint Tests of the Rationality of Both Inflation and Interest Rate Forecasts

A further application of these tests relates to the work of Modigliani and Shiller. Their seminal paper postulates that information on both short-term interest rates and inflation would influence the prices of long-term bonds, along with the proposition that the autoregressive lag structure on the one-period-ahead short rate and inflation forecasts would be "rational" in the sense discussed here. They present evidence supporting this position, yet their evidence is incomplete in two ways. First, they do not actually apply formal statistical tests to the proposition of rationality in the autoregressive lag structures. Secondly, their use of averaged data in the empirical work leads to an aggregation problem that is quite severe.

An example will illustrate this second problem. The Modigliani-Shiller approximation for the expectations hypothesis of the term structure is[24]

$$(10)\qquad RL_t = k + (1-\gamma)\sum_{i=0}^{\infty}\gamma^i r^e_{t+i}$$

where

RL_t = the long bond rate,

k = the liquidity premium,

$\gamma = 1/1+r^*$ where r^* is a representative short rate, and

r^e_{t+i} = the one-period rate expected to hold at $t+i$ conditional on information available at time t.

It implies that if the short rate is a random walk, i.e.,

$$(11)\qquad r^e_{t+1} = r_t$$

then the long rate will be a random walk as well. Working has shown that a variable that has a random walk characterization will, if it is averaged, have an *ARIMA* (0,1,1) time-series process with the correlation coefficient at lag one equal to .25. Hence, if the short rate is a random walk as is the long rate, then averages of both these variables should have the same *ARIMA* (0,1,1) characterization with the 0.25 coefficient at lag one. Using (10) with the averaged short-rate time-series process being the *ARIMA* (0,1,1) described above, the implied time-series process of the averaged long-rate data is not the same as that of the averaged short rate data, as is appropriate. Rather, it will display a time-series process that is closer to that of a random walk. For example, taking the plausible value $\gamma=.95$, the implied time-series

[23] The efficient markets model does not specify whether seasonally adjusted vs. unadjusted data should be used in these tests. Seasonally adjusted data were used in these tests reported in the text because they are more comparable to the rationality tests of the Livingston data found in the literature. However, seasonal adjustment of the *CPI* with the X-11 program tends to "smudge" the data and thus the tests described in the text were repeated with seasonally unadjusted data. The results are similar to those reported in Tables 3 and 4. The likelihood ratio statistic for the 1959:1 to 1969:4 sample period was 23.25 (p-value=.001) and for the 1954:1 to 1976:4 sample period 12.32 (p-value=.055).

[24] The argument here is exactly the same if the more common approximation for a n-period discount bond is used, i.e.,

$$R^n_t = k + \frac{1}{n}\sum_{i=0}^{n-1} r^e_{t+i}$$

where R^n_t = the yield to maturity on the n-period bond.

process derived from (10) of the averaged long-rate series is *ARIMA* (0,1,1) with the autocorrelation at lag one equal to .01 rather than the appropriate .25.[25]

The above example thus indicates that if the data is averaged, equation (10) cannot be used with the lag weights in an autoregressive short-rate equation to derive the lag weights of short rates in a long-rate equation. Modigliani and Shiller's evidence on the rationality of the term structure involves doing exactly this derivation with averaged data, and then comparing these lag weights with those actually estimated from a long-rate equation. Yet as the example here indicates, this is not a valid procedure.

The efficient-markets model discussed in this paper leads to a formal statistical test of the Modigliani-Shiller results discussed above. Including both short-term interest rate and inflation movements as relevant information to the pricing of long-term bonds as is done by Modigliani and Shiller, we can write the efficient-markets model as

$$(12)\quad BRET_t - r_{t-1} = \delta + \alpha_r(r_t - r_t^e) + \alpha_\pi(\pi - \pi_t^e) + \varepsilon_t$$

where r_t = end of period treasury bill rate, and $\pi_t = CPI$ inflation rate over the quarter.

The autoregressive models for r and π are

$$(13)\quad r_t = k_r + \sum_{i=1}^{k} d_i r_{t-i} + \sum_{i=1}^{k} e_i \pi_{t-i} + u_{1t}$$
$$\pi_t = k_\pi + \sum_{i=1}^{k} f_i r_{t-i} + \sum_{i=1}^{k} g_i \pi_{t-i} + u_{2t}$$

and using these autoregressive models to derive expectations

$$(14)\quad BRET_t - r_{t-1} = \delta + \alpha_r\left(r_t - \left(k_r + \sum_{i=1}^{k} d_i r_{t-i} + \sum_{i=1}^{k} e_i \pi_{t-i}\right)\right) + \alpha_\pi\left(\pi_t - \left(k_\pi + \sum_{i=1}^{k} f_i r_{t-i} + \sum_{i=1}^{k} g_i \pi_{t-i}\right)\right) + \varepsilon_t$$

The equations in (13) and (14) can then be estimated jointly as before, and tests of the rationality restrictions can be conducted with the likelihood ratio test. These tests then provide direct information on the Modigliani-Shiller rationality proposition.

These tests and estimates of the efficient-markets model can be found in Tables 5 and 6. The term-structure equation in the MIT-Penn-SSRC quarterly econometric model and the Modigliani and Shiller paper both use a sample period which extends from 1954:4 to 1966:4 and an eighteen-quarter lag on short rates and inflation estimated with a third-order Almon lag. Therefore both the 1954:4 to 1966:4 and the 1954:1 to 1976:4 sample period, as well as the Modigliani-Shiller procedure for estimating the lag structure, are used in the rationality tests conducted here.

The likelihood ratio tests in Table 5 confirm Modigliani and Shiller's results. The restrictions implied by rationality in both the inflation and interest rate forecasts are not rejected at the 5 percent significance level and again the *p*-values are high. Thus Modigliani and Shiller's contention that the term structure of interest rates displays rationality

TABLE 5—MODIGLIANI-SHILLLER TESTS OF FORECAST RATIONALITY

	Sample Period	
	1954:4 to 1966:4	1954:1 to 1976:4
Likelihood Ratio Statistics	13.87	12.90
p-Value	.179	.230

Note: Likelihood ratio statistic is distributed asymptotically as $\chi^2(10)$.

[25] The result is calculated as the *ARIMA*(0,1,1) model for the short-rate average (r^a) with an autocorrelation at lag one of .25, is $\Delta r_t^a = (1 + .268L)u_t$. Then an innovation of $\bar{u}$ would lead to a higher value of r^a by $\bar{u}$ in the initial period and 1.268 $\bar{u}$ thereafter. With $\gamma = .95$, (10) implies that the averaged long rate (RL^a) would be higher by 1.255 $\bar{u}$ initially and 1.268 $\bar{u}$ thereafter. The *ARIMA* model for the averaged long rate would thus be $\Delta RL_t^a = (1 + .011\ L)u_t$, which is an *ARIMA* (0,1,1) with the autocorrelation at lag one equal to .01.

TABLE 6—NON-LINEAR ESTIMATES OF THE MODIGLIANI-SHILLER EFFICIENT MARKETS MODEL

$$BRET_t - r_{t-i} = \delta + \alpha_r\Big(r_t - k_r - \sum_{i=1}^{18} d_i r_{t-i} - \sum_{i=1}^{18} e_i \pi_{t-i}\Big) + \alpha_\pi\Big(\pi_t - k_\pi - \sum_{i=1}^{18} f_i r_{t-i} - \sum_{i=1}^{18} g_i \pi_{t-i}\Big) + \varepsilon'_t$$

$$r_t = k_r + \sum_{i=1}^{18} d_i r_{t-i} + \sum_{i=1}^{18} e_i \pi_{t-i} + u_{1t}$$

$$\pi_t = k_\pi + \sum_{i=1}^{18} f_i r_{t-i} + \sum_{i=1}^{18} g_i \pi_{t-i} + u_{2t}$$

Sample Period 1954:4 to 1966:4

			$\delta = -.0021$ (.0032)	$\alpha_r = -10.9928$ (2.4208)	$\alpha_\pi = -1.8$ (.9752)		
$k_r = .0001$ (.0010)				$k_\pi = .0009$ (.0027)			
$d_1 = .8908$ (.1072)	$d_{10} = .0438$ (.0229)	$e_1 = .0512$ (.1879)	$e_{10} = -.0528$ (.0517)	$f_1 = .1579$ (.0730)	$f_{10} = -.0303$ (.0164)	$g_1 = -.3029$ (.1407)	$g_{10} = -.0618$ (.0352)
$d_2 = -.0640$ (.0743)	$d_{11} = .0431$ (.0228)	$e_2 = .0515$ (.1659)	$e_{11} = -.0456$ (.0478)	$f_2 = .1236$ (.0502)	$f_{11} = -.0215$ (.0170)	$g_2 = .1312$ (.1222)	$g_{11} = -.0744$ (.0322)
$d_3 = -.0372$ (.0463)	$d_{12} = .0401$ (.0247)	$e_3 = .0132$ (.1094)	$e_{12} = -.0373$ (.0524)	$f_3 = .0665$ (.0325)	$f_{12} = -.0123$ (.0187)	$g_3 = .1103$ (.0830)	$g_{12} = -.0800$ (.0344)
$d_4 = -.0147$ (.0308)	$d_{13} = .0351$ (.0271)	$e_4 = -.0158$ (.0897)	$e_{13} = -.0289$ (.0588)	$f_4 = .0237$ (.0232)	$f_{13} = -.0042$ (.0206)	$g_4 = .0859$ (.0657)	$g_{13} = -.0775$ (.0381)
$d_5 = .0038$ (.0267)	$d_{14} = .0284$ (.0287)	$e_5 = -.0366$ (.0903)	$e_{14} = -.0214$ (.0605)	$f_5 = -.0066$ (.0205)	$f_{14} = .0008$ (.0217)	$g_5 = .0593$ (.0621)	$g_{14} = -.0654$ (.0392)
$d_6 = .0184$ (.0277)	$d_{15} = .0201$ (.0288)	$e_6 = -.0502$ (.0921)	$e_{15} = -.0159$ (.0541)	$f_6 = -.0261$ (.0204)	$f_{15} = .0013$ (.0217)	$g_6 = .0317$ (.0614)	$g_{15} = -.0424$ (.0356)
$d_7 = .0295$ (.0282)	$d_{16} = -.0105$ (.0285)	$e_7 = -.0576$ (.0878)	$e_{16} = -.0133$ (.0437)	$f_7 = .0365$ (.0200)	$f_{16} = .0045$ (.0215)	$g_7 = .0045$ (.0581)	$g_{16} = -.0073$ (.0296)
$d_8 = .0372$ (.0270)	$d_{17} = -.0001$ (.0324)	$e_8 = -.0598$ (.0073)	$e_{17} = -.0147$ (.0576)	$f_8 = -.0395$ (.0188)	$f_{17} = -.0188$ (.0248)	$g_8 = -.0211$ (.0514)	$g_{17} = .0413$ (.0374)
$d_9 = .0419$ (.0248)	$d_{18} = -.0114$ (.0468)	$e_9 = -.0579$ (.0636)	$e_{18} = -.0210$ (.1144)	$f_9 = -.0369$ (.0172)	$f_{18} = -.0427$ (.0358)	$g_9 = -.0436$ (.0428)	$g_{18} = .1047$ (.0178)

Sample Period 1954:1 to 1976:4

			$\delta = -.0013$ (.0032)	$\alpha_r = -12.1804$ (1.9664)	$\alpha_\pi = -1.3112$ (.8497)		
$r = .0007$ (.0005)				$k_\pi = -.0016$ (.0013)			
$d_1 = .8013$ (.0841)	$d_{10} = .0326$ (.0178)	$e_1 = .1427$ (.1545)	$\epsilon_{10} = -.0409$ (.0340)	$f_1 = .1277$ (.0563)	$f_{10} = -.0184$ (.0113)	$g_1 = -.0298$ (.1075)	$g_{10} = -.0195$ (.0219)
$d_2 = -.0520$ (.0544)	$d_{11} = .0389$ (.0183)	$e_2 = .0605$ (.1158)	$\epsilon_{11} = -.0406$ (.0266)	$f_2 = .0435$ (.0362)	$f_{11} = -.0040$ (.0115)	$g_2 = .2530$ (.0776)	$g_{11} = -.0233$ (.0186)
$d_3 = -.0410$ (.0328)	$d_{12} = .0431$ (.0192)	$e_3 = .0334$ (.0693)	$\epsilon_{12} = -.0388$ (.0220)	$f_3 = .0051$ (.0215)	$f_{12} = .0106$ (.0121)	$g_3 = .1889$ (.0457)	$g_{12} = -.0228$ (.0180)
$d_4 = -.0296$ (.0208)	$d_{13} = .0449$ (.0197)	$e_4 = .0114$ (.0437)	$\epsilon_{13} = -.0362$ (.0219)	$f_4 = -.0217$ (.0132)	$f_{13} = .0239$ (.0125)	$g_4 = .1348$ (.0281)	$g_{13} = -.0185$ (.0195)
$d_5 = -.0179$ (.0179)	$d_{14} = .0440$ (.0189)	$e_5 = -.0061$ (.0398)	$\epsilon_{14} = -.0331$ (.0241)	$f_5 = -.0383$ (.0113)	$f_{14} = .0345$ (.0121)	$g_5 = .0900$ (.0255)	$g_{14} = -.0113$ (.0210)
$d_6 = -.0063$ (.0190)	$d_{15} = .0401$ (.0173)	$e_6 = -.0195$ (.0448)	$\epsilon_{15} = -.0300$ (.0254)	$f_6 = -.0461$ (.0122)	$f_{15} = .0409$ (.0111)	$g_6 = .0537$ (.0288)	$g_{15} = -.0019$ (.0211)
$d_7 = .0049$ (.0197)	$d_{16} = .0328$ (.0191)	$e_7 = -.0292$ (.0479)	$\epsilon_{16} = -.0273$ (.0254)	$f_7 = .0466$ (.0128)	$f_{16} = .0418$ (.0124)	$g_7 = .0253$ (.0306)	$g_{16} = .0092$ (.0207)
$d_8 = .0153$ (.0192)	$d_{17} = .0218$ (.0301)	$e_8 = -.0357$ (.0466)	$\epsilon_{17} = -.0256$ (.0292)	$f_8 = -.0412$ (.0124)	$f_{17} = .0355$ (.0198)	$g_8 = .0042$ (.0295)	$g_{17} = .0211$ (.0252)
$d_9 = .0247$ (.0183)	$d_{18} = .0069$ (.0509)	$e_9 = -.0394$ (.0415)	$\epsilon_{18} = -.0252$ (.0460)	$f_9 = -.0313$ (.0117)	$f_{18} = .0208$ (.0337)	$g_9 = -.0105$ (.0261)	$g_{18} = .0332$ (.0408)

Note: The $d_2 - d_{18}$, $e_2 - e_{18}$, $f_2 - f_{18}$, *and* $g_2 - g_{18}$ have each been estimated with a third-order polynomial with no fore- or endpoint constraints. Asymptotic standard errors are shown in parentheses.

is supported in these tests, a result we should have expected considering the results of the previous tests in this paper.[26]

III. Conclusions

This paper does provide a response to the question: Are market forecasts rational? The empirical tests conducted here, with one exception, indicate that for the bond market the answer is yes. The bond market data provides no evidence that interest rate forecasts are irrational in this market. Thus, the evidence which finds irrationality in the Goldsmith-Nagan survey of interest rate expectations can be interpreted as casting doubt on the accuracy of this survey measure in describing market expectations.[27] The issue of the accuracy of the Livingston price expectations data, however, is still an open question because irrationality has been found in both the bond market and survey data for the 1959–69 period. Further empirical research on the rationality of this survey data using longer sample periods is needed to help resolve this issue.

This paper has also made the argument that empirical tests in Modigliani and Shiller's seminal paper are incomplete and thus additional empirical tests are required to confirm their conclusion that the term structure of interest rates is "rational." Empirical tests conducted with the methodology outlined in this paper do confirm Modigliani and Shiller's conclusion. This provides further evidence that the bond market does exhibit rational forecasting behavior and is thus efficiently exploiting publicly available information.

Data Appendix

The data sources and definitions of the variables used in this paper are as follows:

$BRET_t$ = Quarterly return from holding a long-term *U.S.* government bond from the beginning to the end of the quarter. The data was obtained from the Center for Research in Security Prices at the University of Chicago, and are described in Lawrence Fisher and James Lorie, and in my 1978 paper.

r_t = The end of quarter ninety-day Treasury bill rate at a quarterly rate. The bill rate data were obtained from the Board of Governors of the Federal Reserve System.

π_t = The *CPI* inflation rate (quarterly rate) calculated from the change in the *log* of the *CPI* (seasonally adjusted) from the last month of the previous quarter to the last month of the current quarter. The *CPI* was collected from *Business Statistics* and *Survey of Current Business.*

[26] Use of seasonally unadjusted *CPI* inflation data rather than the seasonally adjusted data again leads to results similar to those reported in Tables 5 and 6. The likelihood ratio statistic with the unadjusted data for the 1954:4 to 1966:4 period is 17.00 (p-value=.074), and for 1954:1 to 1976:4, it is 11.89 (p-value=.292).

[27] It is possible that the failure to reject rationality in the tests conducted here results from their low statistical power. If the statistical power of the rationality tests using survey data is greater than those in this paper, this might explain the discrepancy between the results in those studies and those here. Yet, there is no obvious reason for this paper's tests to be less powerful than the survey data tests since similar restrictions are tested in both. However, this issue deserves further research attention.

REFERENCES

A. B. Abel and F. S. Mishkin, "A Unified Framework for Testing Rationality, Market Efficiency and the Short-Run Neutrality of Monetary Policy," unpublished paper, Chicago, May 1980.

R. Barro, "Unanticipated Money Growth and Unemployment in the United States," *Amer. Econ. Rev.*, Mar. 1977, *67*, 101–15.

———, "Unanticipated Money, Output, and the Price Level in the United States," *J. Polit. Econ.*, Aug. 1978, *86*, 549–80.

J. A. Carlson, "A Study of Price Forecasts," *Annals Econ. Soc. Measure.*, Winter 1977, *6*, 27–56.

Eugene F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, *25*, 383–417.

———, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

———, (1976a) *Foundations of Finance*, New York 1976.

______, (1976b) "Forward Rates as Predictors of Future Spot Rates" *J. Finan. Econ.*, Oct. 1976, *3*, 361–77.

______ **and G. W. Schwert,** "Asset Returns and Inflation," *J. Finan. Econ.*, Nov. 1977, *5*, 115–46.

Lawrence Fisher and James H. Lorie, *A Half Century of Returns on Stocks and Bonds Securities, 1926–1976,* Chicago 1977.

J. A. Frenkel, "The Forward Exchange Rate Expectations, and the Demand for Money: The German Hyperinflation" *Amer. Econ. Rev.*, Sept. 1977, *67*, 653–70.

B. Friedman, "Survey Evidence on the 'Rationality' of Interest Rate Expectations," *J. Monet. Econ.*, Oct. 1980, *6*, 153–66.

H. Glesjer "A New Test for Heteroscedasticity," *J. Amer. Statist. Assn.*, Mar. 1969, *64*, 316–23.

Stephen M. Goldfeld and Richard F. Quandt, *Non-Linear Methods in Econometrics*, Amsterdam 1972.

______ **and** ______, "Some Tests for Homoscedasticity," *J. Amer. Statist. Assn.*, June 1965, *60*, 539–47.

C. W. J. Granger and Paul Newbold, "Spurious Regression in Econometrics," *J. Econometrics*, July 1974, *2*, 111–120.

S. J. Grossman and J. E. Stiglitz, "Information and Competitive Price Systems," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 246–53.

Joseph A. Livingston, surveys published twice yearly, *Philadelphia Sunday Bulletin*, 1948–71; *Philadelphia Inquirer*, 1972 on.

F. S. Mishkin, "Efficient-Markets Theory: Implications for Monetary Policy" *Brookings Papers, Washington* 1978, *3*, 707–52.

______, "Monetary Policy and Long-Term Interest Rates: An Efficient Markets Approach," *J. Monet. Econ.*, Jan. 1981, *7*, 29–56.

F. Modigliani and R. J. Shiller, "Inflation, Rational Expectations, and the Term Structure of Interest Rates," *Economica*, Feb. 1973, *40*, 12–43.

D. J. Mullineaux, "On Testing for Rationality: Another Look at the Livingston Price Expectations Data," *J. Polit. Econ.*, Apr. 1978, *86*, 329–36.

J. F. Muth, "Rational Expectations and the theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

C. R. Nelson and G. W. Schwert, "Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis that the Real Rate is Constant," *Amer. Econ. Rev.*, June 1977, *67*, 478–86.

J. E. Pesando, "A Note on the Rationality of the Livingston Price Expectations," *J. Polit. Econ.*, Aug. 1975, *83*, 849–58.

T. J. Sargent, "A Note on Maximum Likelihood Estimation of the Rational Expectations Model of the Term Structure" *J. Monet. Econ.*, Jan. 1979, *5*, 133–44.

R. J. Shiller, "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure," *J. Polit. Econ.*, Dec. 1979, *87*, 1190–219.

H. Working, "Note on the Correlation of First Differences of Averages in a Random Chain," *Econometrica*, Oct. 1960, *28*, 916–18.

Goldsmith-Nagan Bond and Money Market Letter, Washington, various issues.

U.S. Office of Business Economics, *Surv. Curr. Bus.*, Washington, various issues.

On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence

By ISAAC EHRLICH*

While classical economists generally considered deterrence of potential offenders the sole function of criminal sanctions and the principal instrument of crime control, the emphasis in modern criminological thought has shifted from deterrence toward the rehabilitation and incapacitation of convicted offenders.[1] The emphasis on direct control of the behavior of identified offenders, occasionally mislabeled "specific deterrence," reflects, in part, the growing interest among modern criminologists in individual causes of crime as well as in offenders as individuals: The promise of successful rehabilitation and control of known offenders, many of whom are poor and uneducated, has a strong humanitarian and moral appeal. It has implications for the behavior and future income, if not the actual welfare, of these individuals. The direct control of individual offenders has been conceived of, however, not just as a means of providing private relief, but as an effective check on the total incidence of crime. The restraining, retraining, counseling, and direct guidance offered to convicted offenders have been viewed as forms of social engineering aimed at effecting a reallocation of human resources away from crime toward socially more useful endeavors.

Several evaluation studies have been conducted in recent years to assess the effectiveness of rehabilitation and other methods of individual control as means of crime prevention.[2] All seem to share a similar methodological approach in that they attempt to estimate or predict analytically the impact of these methods on individual recidivism (i.e., the rate of offenders' reentry into crime). They then implicitly equate the observed or anticipated outcomes at the individual level with those in the aggregate. The point of departure of this paper is the distinction between effectiveness of means of crime control at the aggregate or *market* level as opposed to the individual level. If the flow of offenses of specific types reflects, by and large, not the capricious outcome of biological or social idiosyncracies, but the equilibrating interplay of systematic "supply and demand" forces, then the effectiveness of individual control programs must be evaluated not by their anticipated initial effect on the supply of offenders, but by their ultimate effect on the equilibrium volume of offenses. Indeed, recognition of the existence and the role of equilibrium in the "market for offenses" is shown to lead to important modifications in previous conclusions concerning the relative efficacy and efficiency not only of methods of control of individual offenders, but of means of deterrence as well.

The plan of the paper is as follows: the general components of the market for offenses are presented in Section I, and the basic equilibrium analysis concerning the effectiveness of public intervention in that market is developed in Section II. Sections III and IV present more specific implications concerning crime control via rehabilitation and incapacitation and examine some related empirical evidence. In Section V the analysis is used to derive additional implications for

*Professor of economics, State University of New York-Buffalo. I have benefited from stimulating discussions with Aaron Director, and from comments by the participants of the sixth Interlaken Seminar on Analysis and Ideology, Interlaken, Switzerland. This research was partially supported under grant no. 79-N1-AX-0040 from the National Institute of Law Enforcement and Criminal Justice. Points of view are my own and do not necessarily reflect the position of the U.S. Department of Justice.

[1] See Leon Radzinowitz, chs. 2 and 3.

[2] See the references provided in Sections III and IV of this paper.

the choice of optimal criminal sanctions. A number of general implications concerning the treatment of individual offenders and specific types of crime are then illustrated in Section VI.

I. The Market for Offenses

Essential to a comprehensive economic model of crime is the assumption that potential offenders, victims, buyers of illegal goods and services, and the law enforcement authorities all behave according to the fundamental rules of maximizing behavior. It is further postulated that the activities of all agents are coordinated and made mutually consistent at the market level through the effects of explicit or implicit prices (see my 1979 paper). In previous works (for example, the seminal paper by Gary Becker), equilibrium in the market for offenses has been synthesized out of the interplay between only two identified groups: potential offenders, representing the "supply" side of the market; and law enforcement authorities, representing public intervention. Missing in these formulations was a systematic consideration of the roles of potential victims and buyers of illegal goods and services who, by their respective demand for safety or for illegal transactions, dictate the shape of the private "derived demand" for offenses. As is the case in analyses of displacements of equilibria in legitimate markets, a rigorous examination of the effectiveness of public intervention in the market for offenses requires an explicit consideration of both private supply and demand forces in determining the equilibrium volume of offenses at any given level of public intervention.

In what follows I shall first introduce the basic components of such a broader, and more relevant, market system, with an emphasis on those segments of the system that have not been adequately considered before.

A. *Supply of Offenses*

An elaborate analysis of the supply side of the market appears in earlier studies (see, for example, my 1974 paper). For the sake of a simple diagrammatical exposition, and without affecting the generality of the subsequent equilibrium analysis, I shall present here a simple version of the model in which attitudes toward risk are assumed to be neutral. The offender's supply of offenses of any given type $s(\pi)$ is then expected to be, in general, a nondecreasing function of his expected net return per offense $\pi=d-pf$, where d denotes his differential payoff from the illegitimate over an alternative (say, legitimate) activity, net of all the direct costs involved in carrying out the offense,[3] and pf denotes his expected direct or opportunity cost due to the criminal sanction imposed (f), with p denoting the probability of apprehension and punishment. Formally, $s'(\pi)\geqslant 0$.[4]

To further simplify the analysis of aggregation of individual supply functions, let the net return per offense be identical to all offenders. Then the aggregate frequency of offenses, $q=S(\pi)$, likewise would be a nondecreasing function of the net return per offense. The proof of the latter proposition follows from the presumed existence of a continuous distribution of individual preferences for participation in illegitimate activities. The latter can be represented by a density function of critical entry returns which are sufficient to induce different individuals' entry to the market for offenses, $\gamma(\pi^*)$.[5]

[3] More specifically, $d=d(w_i, w_l, c)$ is a function of the gross payoff per offense w_i, net of the various expected costs of "producing" the offense which depend, in turn, on the effective measures of self-protection by the victim c, and the foregone value of the offenders' time in an alternative (legitimate) activity w_l. In crimes committed for strictly nonpecuniary objectives, $w_i=0$, and therefore both d and π are negative quantities. The supply of offenses can still be depicted, however, as an increasing function of d or $(-pf)$.

[4] The prediction that individual participation in criminal activity is a nondecreasing function of the monetary net return π would in general hold unambiguously, of course, only for compensated changes in π that left the offender's relevant real income unchanged. It holds generally, however, in connection with an offender's incentive to first enter the criminal market, and also for the choice to intensify illegal activity on the assumption that the income effect of a change in any of the relevant components of π, whatever its direction, is not sufficiently strong to offset the corresponding substitution effects on time allocation in favor of crime.

[5] More generally, if individuals faced identical criminal payoffs but differed with respect to the legitimate

Let the corresponding density of additional offenses supplied at these critical net returns be given by $g(\pi^*)$. Clearly, $g(\pi^*)$ would be a continuous and positive function even if $s'(\pi)=0$. Since the mean supply-of-offenses function is the cumulative density function $S(\pi)=\int_{-\infty}^{\pi} g(\pi^*)\,d\pi^*$, it would then be necessarily nondecreasing in π. In general, the more condensed the frequency distribution of critical entry returns about particular values of π, the more elastic will be the aggregate supply of offenses about these values. Only in the case where offenders constituted a "noncompeting group" totally irresponsive to incentives would the aggregate supply of offenses schedule be completely inelastic at a fixed supply of offenses.

B. *The Private Derived Demand for Offenses*

The concept "demand for offenses" may, on first glance, seem paradoxical in reference to those offenses which impose negative externalities on all relevant parties. Some criminal activities, especially those labeled "victimless crimes," do take place, however, under the patronizing influence of second parties. There are, in fact, explicit markets for voluntary exchanges in all illicit goods and services, including goods that are acquired through the commission of crimes against property and person. The willful, direct or derived, demand for offenses—whether desired for their own sake or as a means of satisfying the demand for stolen goods—forms at least one fragment of the private demand for these offenses, which is expected to obey all the fundamental laws of demand theory.[6]

There does exist an implicit private demand schedule for offenses of all types, however, including those that harm second parties. Such a schedule is *derived* from the private demand for safety. As a formal construct, the demand schedule for offenses represents the average potential payoff per offense at alternative frequencies of offenses $d(q)$. Measured as the differential value of the loot (if any) over the direct and opportunity cost of "production" incurred by the offender, the potential payoff d is in large measure a function of the level of vulnerable nonhuman and human assets possessed by potential victims. In addition, it is dictated by the amount of self-protection and self-insurance (c) they provide to protect these assets.[7] Burglar alarm systems, guards, locks, safe deposit boxes, selected places of residence, and restricted travel all serve the similar purpose of decreasing the gross loot per offense, or increasing the cost and effort to the offender of committing the offense. Optimal expenditure on protection, especially if combined with an optimal purchase of market insurance at actuarially fair terms, can be shown in turn to be a continuous and increasing function of the rate of offenses (i.e., the objective risk of victimization), or $c'(q)\geqslant 0$.[8] Since the potential payoff per offense is a decreasing function of private protection, all other determinants of illegitimate and legitimate opportunities held constant, it would therefore be a decreasing function of the crime rate itself.

wages available to them, $\gamma(\pi^*)$ would be determined by the joint-probability distribution of individuals' preferences for crime and their alternative earning opportunities.

[6] For an attempt to implement this idea empirically in a study of auto theft, see Walter Vandaele.

[7] Self-protection by victims may also contribute to the probability that an offender is apprehended and punished, which would further reduce his expected net return π. For simplicity, this source of interaction between private and public protection will not be considered here.

[8] This result can be proved unambiguously for the decision to provide self-insurance (activities which reduce the potential size of the loot), and even self-protection (activities which reduce the personal risk of victimization), provided that protection is combined with full market insurance, and that the marginal benefits from protection increase as the average risk of victimization q (the general crime rate) rises. (For a general analysis see my paper with Gary Becker.) Indeed, since expected income is assumed to be continuous, differentiable, and *strictly concave* in the real outlays on self-insurance and protection, c, the theorem of the maximum (see, for example, Hal Varian) guarantees that the optimal expenditure on protection per capita $c^*(q)$ also will be a continuous, differentiable, and increasing function of q. Furthermore, by assumption, $d=d(w_i, w_l, c)$ is a continuous and decreasing function of c. Thus, d itself is expected to be a continuous and decreasing (indirect) function of q.

Finally, the shape of the market derived demand for offenses involving material gains must also exhibit "diminishing marginal returns" from the stock of available targets. With opportunities for gains from property crimes unevenly distributed in the population, optimal selection of criminal targets by cost-minimizing offenders implies that, as the aggregate volume of offenses increases, the marginal targets selected would be associated with greater costs of production per dollar gained. For this reason alone, the differential return per offense is expected to be a decreasing function of the aggregate frequency of offenses.

Since all the components of the private derived demand for offenses are expected to be negatively related to the frequency of offenses (all prices of protective devices, wealth, legitimate earning, and productivity parameters held constant), their vertical sum is also expected to be a continuous and decreasing function of the rate of offenses. Formally, $d=D(q)$ with $D'(q)\leqslant 0$.

C. *Public Enforcement*

Since criminal activities by definition create external diseconomies, and since private self-protection or private enforcement of criminal laws are themselves associated with various externalities and some properties of a nonexclusionary public good, there is a generally recognized incentive for public intervention in the market for offenses.

If social optimality is presumed to be founded on the principle of maximizing the efficiency of overall resource allocation rather than on any measure of vengence, moral "justice," or equality in the distributive outcomes of law enforcement, then the target of public enforcement and protection can be stated in terms of minimizing the aggregate, or per capita losses from crime. The relevant social loss function in this formulation is generally comprised of three principal cost functions: the direct social damage from offenses Δ, defined as the overall loss to society from crime over the private gains to offenders; the *direct* cost of interference in the market for offenses C; and the social costs resulting from the subsequent treatment imposed on convicted offenders. Note that, since criminals cannot collect as earnings all the damage they impose on victims (for example, the value of life and property destroyed, the real cost of insurance and protection against victimization, and the value of resources used to commit offenses), the net social damage from crime Δ is expected to be positive in connection with "simple" theft and fraud, as well as serious felonies. For the case of intervention via conventional law enforcement activity (see Becker), C summarizes the costs of apprehending and convicting offenders, and the costs of subsequent treatment are those resulting from imposition of criminal sanctions. Formally,

$$(1)\quad L(q)=\Delta(q)+C(q,p)+b(t)pfq$$

where p and f denote the probability and severity of the specific criminal sanctions imposed, and b is a "social cost multiplier" which transforms the private cost of punishment to the offender into social cost terms, and which depends on the form of punishment used (t).

Equation (1) represents a public welfare criterion for determining an optimal policy of crime control. To determine the expected criminal sanction pf, for example, equation (1) must be minimized subject to a) the crime-response function, summarizing the effectiveness of the sanction (whatever its form) in reducing the actual volume of offenses, and b) the production function of direct law enforcement activity which determines the properties of $C(q, p)$.

Numerous behavioral propositions emanating from this formulation have been discussed at length by Becker and in my 1977 paper. One implication that is of particular relevance here concerns the optimal social response to changes in the frequency of offenses due to specific exogenous factors. Note, first, that such societal response function does not constitute an independent derived-demand-for-offenses schedule, since public enforcement is a monopolized state activity which is therefore dependent on both the private supply and demand schedules. Rather, the law enforcement authority sets its optimal enforcement policy after taking account of the parameters (elasticities) of these schedules, as well as the parameters controlling all cost terms included in equation (1). Given the vector of these parameters ϕ, however, it can be shown under fairly general conditions that the optimal values of the enforcement instrument pf will be adjusted upward whenever the frequency of offenses rises due to changes in other exogenous or random factors.[9] This pattern of societal response helps guarantee, of course, the stability of equilibrium in the general market for offenses.

The market system introduced in the preceding discussion can be illustrated most simply by assuming that public intervention assumes the form of enforcement of purely deterring sanctions such as fines. In that case, law enforcement does not affect the private demand and supply relationships directly, but operates like an excise tax or tariff in the amount of $\tau = pf$. Equilibrium in the market for offenses would then be the solution of the simplified system[10]

$$(2) \qquad q^s = S(\pi) \qquad \text{with } S'(\pi) > 0$$

$$(3) \qquad d = D(q^d) \qquad \text{with } D'(q) \leqslant 0$$

$$(4) \qquad pf = \tau(q^d | \phi) \qquad \text{with } \tau'(q) \geqslant 0$$

$$(5) \qquad \pi \equiv d - pf$$

$$(6) \qquad q^s = q^d$$

The supply, demand, and "tax" schedules given in equations (2)–(4) are depicted graphically in Figure 1, with q^s and q^d denoting the quantities of offenses "supplied" and "demanded," respectively. Equation (5) expresses the necessary condition for equilibrium, with the solution at point Q in Figure 1 seen to be stable by virtue of the properties of equations (2)–(4).

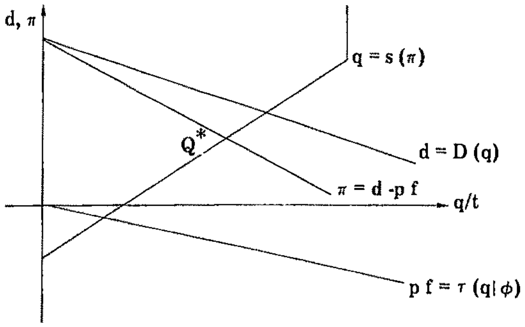


FIGURE 1

II. Individual Control and Deterrence: An Equilibrium Analysis

The three basic measures of crime control most frequently discussed in the criminological literature are deterrence, incapacitation, and rehabilitation. Deterrence essentially aims at modifying the "price of crime" for all offenders, potential and actual. It is analogous to any method of public intervention that seeks to modify the market net return from crime, π, through either punishment, an improvement in employment opportunities in legitimate labor markets (hence a reduction in d), or related efforts. Rehabilitation and incapacitation, in contrast, seek to remove a subset of convicted offenders from the market for offenses either by relocating them in legitimate labor markets, or by excluding them from the social scene for prescribed periods of time. Typical means of incapacitation such as imprisonment exert, of course, both incapacitative and deterrent effects. For obvious methodological purposes, however, the term incapacitation will be used in this analysis to convey the distinct role of physical removal as means of reducing individual recidivism at any given level of net punishment.

[9] This result, subject to assumptions required for fulfillment of second-order conditions, is discussed in my paper with Joel Gibbons in Proposition 1, p. 49.

[10] The simplified market system expressed in equations (3) and (4) abstracts from any *direct* (technical) interdependencies between private protection and public enforcement. Thus, a change in τ is not assumed to cause a shift in the private demand schedule $D(q)$. Note, however, that equations (3.) and (4) do express *indirect* interdependence between private and public protection because the former is shown to be an increasing function of $q(D'(q) \leqslant 0)$. Thus, for example, a decrease in public law enforcement due to an exogenous factor (say, a police strike) is clearly expected to bring about an increase in the amount of private self-protection provided, because of the resulting increase in the crime rate.

The effectiveness of rehabilitation and incapacitation in curbing individual recidivism and their overall potential quantitative significance will be considered in subsequent sections. In this section, I shall try to determine the efficacy of individual control methods relative to deterrence at the market level under any assumed level of their efficacy at the individual level. Effective control of convicted offenders with positive probabilities of recidivism would then amount, in a steady state of the market, to a leftward shift in the aggregate supply-of-offenses schedule from the initial equilibrium position. For illustration, suppose that the market supply curve had the linear shape of the cumulative uniform probability distribution of entry net returns, π^*, sufficient to induce members of the population at large to enter the market for offenses, all individual supply curves being inelastic about an arbitrary number of offenses. Let the subset of individual offenders apprehended and removed from the market be randomly drawn from the full set of active offenders at the initial equilibrium net return, π_0. Then the initial supply curves S_0S_0 would be reshaped into $S_0A_0S_1$ if individual control were via incapacitation, or into $S_0A_0S_2S_0$ if control were via rehabilitation, provided that the latter resulted in an equal absolute increase in the entry net returns the subset of rehabilitated offenders would now require before reentering the market for offenses. As Figure 2 shows, the change in the equilibrium frequency of offenses can differ drastically from the potential aggregate change corresponding to the total number of offenders removed under incapacitation or rehabilitation. The difference between the actual and anticipated effects of individual control methods is indicated by the distance A_0A_1 in the case of incapacitation and by the distance A_0A_2 in the case of rehabilitation.

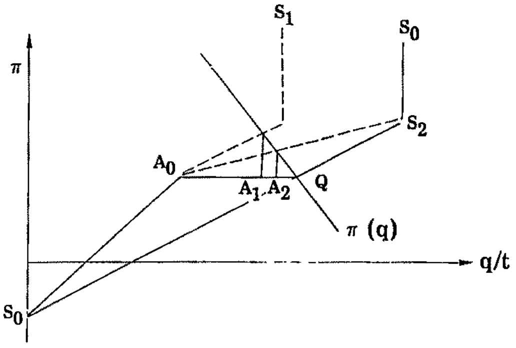


FIGURE 2

The source of the difference between any successful rehabilitative or incapacitative effects at the individual and the market levels is the replacement of individual offenders who are successfully removed from the market for offenses by veteran offenders or new entrants who are induced by the prevailing opportunities for illegitimate rewards to fill the vacancies created by the departing offenders. Since control of the behavior of convicted offenders per se does not involve changes in expected criminal sanctions or in the private derived-demand-for-offenses schedule, the departure of individual offenders and the accompanying reduction in the frequency of offenses will temporarily increase the market net return from offenses. The increased rewards due, say, to higher demand prices for illegal goods or to a lower amount of private protection, in turn would operate as a signal to potential participants to enter or reenter the market, as the case may be, and would induce active offenders to adjust their participation in illegitimate activities upward. This replacement effect, offsetting the initial removal effect exerted by methods of individual control, would be inevitable as long as supply elasticities were greater than zero, private demand elasticities less than infinite, and alternative law enforcement activities unchanged.

These conclusions can be expressed more rigorously through the following formal analysis. Assume, for convenience, that individual supply-of-offenses schedules were all of a constant elasticity variety, differing only in individual constant terms. Then the mean supply-of-offenses function would have the same constant elasticity as the individual functions regardless of the mix of offenders operating.[11] Similarly, assume that the implicit market demand schedule for offenses,

[11] That is, if $q_j = A_j\pi^\alpha$, all j, then

$$\frac{1}{N}\sum_j q_j = \frac{1}{N}A_j\pi^\alpha = q$$

incorporating a fixed expected sanction imposed through public law enforcement, is also of a constant elasticity variety. Market equilibrium would then be the solution of the three equation system

$$q^s = A_0 \pi^\alpha \tag{7}$$

$$q^d = B_0 \pi^{-\beta} \tag{8}$$

$$q^d = q^s \tag{9}$$

If individual control methods could effectively reduce the recidivism rate of controlled offenders, effective control would be tantamount, in the context of the present model, to a finite reduction in the value of the coefficient A_0 in equation (7) with no change in the coefficients α, β, or B_0. Let the percentage change in A_0 corresponding to a given program of individual control (J) be given by $\tilde{A}_0 = \partial \ln A_0/\partial J$. The effect of the program on the equilibrium frequency of offenses will then be given by

$$-\frac{\partial \ln q^*}{\partial J} = \frac{\beta}{\alpha+\beta}\tilde{A}_0 \tag{10}$$

where q^* is the solution to equations (7), (8), and (9). The term $\beta/(\alpha+\beta)$ indicates the extent to which the removal effect is offset by the equilibrium replacement effect. Clearly,

(11)

$$\frac{\beta}{\alpha+\beta} = \begin{cases} 0 & \text{if } \alpha = \infty \quad \text{or } \beta = 0 \\ 1 & \text{if } \alpha = 0 \quad \text{or} \rightarrow 0 \text{ as } \beta \rightarrow \infty \\ <1 & \text{if } \alpha > 0, \quad 0 < \beta < \infty \end{cases}$$

The replacement effect would then be complete if the supply-of-offenses schedule was infinitely elastic and the market demand elasticity was zero about the initial equilibrium position. It would be nil only if the supply elasticity was zero and the demand elasticity was infinite. In all other cases the actual efficacy of individual control programs would be moderated by the multiplier $\beta/(\alpha+\beta)$.

And what about deterrence? If the extent of public control via law enforcement activity were confined to setting the expected punitive tax, or fine pf, with no direct control over the parameters of the differential payoff schedule $d = D(q)$, then, in the context of the present model, the effect of such control would be tantamount to a reduction only in the initial equilibrium rate of criminal return π^0 with no change in α, β, or A_0. The effect of a percentage change in π^0 via the deterrence program τ on the equilibrium rate of offenses is then given by[12]

$$-\frac{\partial \ln q^*}{\partial \tau} = \frac{\beta\alpha}{\alpha+\beta}\tilde{\pi}^0 \tag{12}$$

where $\tilde{\pi}^0 \equiv -\partial \ln \pi^0/\partial\tau$. Clearly then,

$$\frac{\beta\alpha}{\alpha+\beta} \begin{cases} \rightarrow \alpha & \text{as } \beta \rightarrow \alpha \\ \rightarrow \beta & \text{as } \alpha \rightarrow \infty \\ <\alpha \geqslant 0 & \text{if } 0 < \alpha \leq \infty, \beta \geqslant 0 \end{cases} \tag{13}$$

Unlike the efficacy of methods of individual control in reducing the aggregate rate of offenses, which is shown in equation (11) to be a decreasing function of the elasticity of the market supply-of-offenses schedule α, the efficacy of general deterrence is an increasing function of the latter elasticity essentially because deterrence operates like a change in the initial market price. The equilibrium analysis further indicates, however, that the efficacy of both deterrence and methods of individual control of offenders are increasing functions of the elasticity of the market demand curve for offenses.

or $q = A_0\pi^\alpha$ where $A_0 = \frac{1}{N}\sum A_j$

and N denotes the population at large.

[12] By equation (8), given the initial market equilibrium $q = q^0$ and $\pi = \pi^0$, $B_0 = (\pi^0)^\beta q^0$. The equilibrium frequency of offenses, as solved from equations (7)–(9) is given by

$$q^* = A_0^{\beta/(\alpha+\beta)} B_0^{\alpha/(\alpha+\beta)}$$

Since the deterrence program affects directly π^0, rather than B_0, the effect of a change in π^0 on q^* is then easily found to be

$$\frac{\partial \ln q^*}{\partial \ln \pi^0} = \frac{\beta\alpha}{\alpha+\beta}$$

III. To What Extent Can Rehabilitation Reduce Crime?

The term rehabilitation, as used in connection with criminal behavior, has come to denote various methods of treatment of convicted offenders aimed at reducing individual recidivism through imposition of specific positive incentives. In the last few decades a variety of rehabilitative methods ranging from therapy to vocational training programs have been tried in the United States and elsewhere, in some cases on a significant scale.[13]

The effectiveness of these programs has been assessed exclusively in terms of their success at the individual level. The evidence for effectiveness even at that level has been rather meager. Numerous studies indicate little success, if not outright failure, of most programs in bringing about any enduring rehabilitative outcomes for treated offenders.

It is possible that the degree of actual success has been greater than what many studies estimate, or that the incentives provided through training and related programs have been insufficient. Pursuing the economic approach developed here, one can say only that while successful rehabilitation may be quite costly to achieve, it is in principle a function of the quantity and quality of resources invested in, or the implicit subsidy provided to, individual offenders toward acquisition of various legitimate skills. What has been entirely missing in studies of the rehabilitation experience is the realization that whatever its effect on treated offenders, its role at the market level would be hampered by three additional constraints: a) the typically small proportion of the potential offender population that can be subjected to treatment; b) equilibrating forces at the market level;[14] and c) counterincentives arising from benefits conferred on convicted offenders. These considerations can be spelled out formally as follows.

Denote the total number of persons wishing to enter and participate in the market for offenses at a given rate of criminal returns π_0 by $S^e(\pi_0)$, and assume that on average offenders commit $k(\pi_0)$ offenses per period. Each period a fraction p of all participants in the illegal market is apprehended and convicted, and a fraction r of these offenders is ultimately rehabilitated after fully serving the criminal sanctions imposed. Successful rehabilitation implies in turn that, on average, rehabilitated offenders are removed from the market for offenses for L periods, where L may coincide, at most, with the offender's remaining labor market horizon. The market for offenses is assumed to be free of secular growth.[15]

Under these assumptions it can be shown that, as long as π_0 remained unchanged, an increase of 1 percent in the fraction of successfully rehabilitated convicts would lead in a steady state to a decline in the frequency of offenses committed in the amount

$$(14)\qquad \sigma_r^{max} = -\frac{\partial \, ln\, q(\pi_0)}{\partial \, ln\, r} = \frac{rpL}{1+rpL}$$

where $-\partial \, ln\, q(\pi_0)/\partial \, ln\, r$ is analogous to the term A_0 in equation (10). Clearly, in view of the small magnitude of r in practice, σ_r^{max} is expected to be quite small.

Moreover, the equilibrium flow of offenses would be modified by the change in the market net return from crime accompanying the process of rehabilitation: In terms of the simplified market model given by equations (7)–(9), if rehabilitation does not affect any of the parameters of the supply and demand schedules other than A_0, then, by equation (10),[16]

$$(15)\qquad \frac{\partial \, ln\, q^*}{\partial \, ln\, r}\Bigg|_{\substack{B_0,\alpha,\beta.\\ pf\; constant}} = \frac{\partial \, ln\, A_0}{\partial \, ln\, r}\,\frac{\beta}{\alpha+\beta} = \frac{rpL}{1+rpL}\cdot\frac{\beta}{\alpha+\beta}$$

[13] An extensive survey of these programs is included in Douglas Lipton, Robert Martinson, and Judith Wilks. See also Leslie Wilkins, James Robison and Gerald Smith; Roger Hood; Walter Baily; Phillip Cook; James Wilson.

[14] The following analysis is based on the implicit assumption that the increased supply of graduates of rehabilitation programs in the legitimate sector of the economy is sufficiently small in relative terms so as not to cause any perceptible reduction in legitimate wages available to offenders.

[15] But see the analysis in the Appendix. Positive population growth is shown to reduce the magnitude of σ_r^{max} in equation (14).

[16] Note that the process of replacement due to the vacancies created by successfully rehabilitated offenders would be operative not only at the market level, but

where $\beta/(\alpha+\beta)$ expresses the replacement effect at the market level.

But the effect of rehabilitation on the equilibrium frequency of offenses is even more complex. The reason is that successful rehabilitation confers an implicit *subsidy* on potential offenders by offering training and employment benefits at public expense. Even if the rehabilitation programs were not carried out at the expense of the criminal sanctions, but rather in addition to them, the provision of rehabilitative net benefits—to the extent that they are positive—necessarily enhances the anticipated net return from crime to the potential offender (π_0) by the magnitude of the rehabilitation subsidy per offense, g. Put differently, the rehabilitation benefits provided to actual offenders *ex post* produce a counterdeterrent effect on potential offenders *ex ante*. By equation (12) the total effect of the rehabilitation subsidy is given by

$$(16)\quad \varepsilon_r \equiv -\frac{\partial \ln q^*}{\partial g}\bigg|_{\substack{\alpha,\beta\\ pf\, constant}} = \frac{\beta}{\alpha+\beta} \times \left[\frac{\partial \ln r}{\partial g}\,\frac{rpL}{1+rpL} - \alpha\frac{\partial \ln \pi_0}{\partial g}\right]$$

Clearly, if the subsidization effect $(\partial \ln \pi_0/\partial g) \equiv \tilde{\pi}$ were sufficiently high, rehabilitation may increase rather than lower the actual frequency of crime in the population.[17] If the rehabilitative subsidy were negative, however ($g<0$), its effect on crime would be analogous to that of a criminal sanction.

IV. The Preventive Effect of Imprisonment: Deterrence or Incapacitation?

Imprisonment produces an incapacitative as well as a deterrent effect. The argument that it derives its efficacy and efficiency mainly from its incapacitative value[18] is deficient, however, on several important grounds. First, although imprisonment temporarily eliminates participation in criminal activity outside of prison walls, it does not stop it inside. Moreover, since imprisonment is likely to result in the relative depreciation of legitimate knowledge and skills, it may lead to an increase in the rate of recidivism of imprisoned offenders in their postrelease period. Part of the incapacitative value of temporary incarceration, then, may be offset by its "hardening" effect on the same offenders.

Even if hardening effects are ignored, the maximum incapacitative value of imprisonment, measured in elasticity terms, can be found to be rather modest in practice in view of the small magnitudes of both the probability that a potential offender at large be apprehended and imprisoned p, and the typical length of his actual incarceration T. By a direct application of the preceding analysis of the removal effect associated with rehabilitation (see the Appendix) a 1 percent change in either p or T would generate a maximum incapacitative effect on the crime rate equal to

$$(17)\quad \sigma_I^{max} \equiv -\frac{\partial \ln q(\pi_0)}{\partial \ln pT} = \frac{pT}{1+pT}$$

As the analysis in the following section will show, it is the elasticity of the crime rate with respect to incapacitative measures, rather than the absolute "marginal product" of the latter, which determines the marginal social "revenue" from the allocation of resources to

among offenders participating in the rehabilitation program as well. In particular, the knowledge of illegitimate openings created by rehabilitated offenders, and sometimes cooperation among offenders in the treatment group, would induce greater participation in illegitimate activity on the part of those in the group with greater comparative advantage in crime. This is one reason why the average rate of recidivism among graduates of rehabilitation programs may not differ markedly from that among offenders outside the program, as some evaluation studies report (see fn. 13).

[17] Again, this counterdeterrent effect may operate at the individual as well as at the market level, as long as recidivism on the part of graduates of rehabilitation programs would not foreclose their opportunities for obtaining future rehabilitation benefits. Since offenders undergoing rehabilitation are in a position to assess the rehabilitation benefits with greater certainty than other offenders, their average rate of recidivism may even rise relative to that of other convicts.

[18] See, for example, Shlomo Shinaar and Reuel Shinaar, and Jan Chaiken, Michael Lawless, and Keith Stevenson.

the production of incapacitative instruments. Thus, the argument that the marginal social value of incapacitation is high because the majority of offenses are committed by a small number of offenders (i.e., that the number of offenses committed by a typical offender k is high), turns out to be irrelevant for the determination of optimal policy vis-à-vis incapacitative instruments on the margin, because k itself does not influence the magnitude of σ_I^{max}.

Again, the maximal incapacitative effect of imprisonment is necessarily an overstatement of the actual effect because it fails to account for the replacement effect caused by the displacement in market equilibrium. However, changes in p and T produce shifts in both the market demand-for-offenses schedule (reflecting deterrence) and in the supply-of-offenses schedule (reflecting incapacitation). In the Appendix it is shown that the steady-state supply-of-offenses schedule can be written generally as

$$(18) \qquad q = A_0(pT)S(\pi)$$

where, in a steady state with no growth, $A_0 = 1/(1+pT)$. Thus, in terms of the market model given in equations (7)–(9), the total effect of equal percentage changes in either p, T, or their product pT on the equilibrium rate of offenses is given by

$$(19) \quad \varepsilon_p = \varepsilon_T \equiv -\left.\frac{\partial \ln q^*}{\partial \ln pT}\right|_{\alpha,\beta\ constant}$$

$$= \frac{\beta}{\alpha+\beta}\frac{pT}{1+pT} + \frac{\beta\alpha}{\alpha+\beta}\frac{\partial \ln \pi_0}{\partial \ln pT}$$

Equation (19) points to a serious overstatement of the pure incapacitative effect of imprisonment, or its share in the total preventive effect of imprisonment as addressed analytically and empirically in many recent studies, because of failure to assess incapacitation effects within the context of equilibrium analysis.[19] As equation (19) shows, the higher the elasticity of the supply schedule α, the lower the actual incapacitative effect of imprisonment, and the higher the fraction of the overall preventive effect of imprisonment attributable to deterrence. Although the potential hardening effect of incapacitation that counteracts its potential removal effect has not been added as a third argument in equation (19), it is apparent that both the incapacitation and the hardening effects of imprisonment would be minor in practice if α were sufficiently high.

An illustration of the potential empirical importance of the incapacitative effect of imprisonment can be provided by evaluating the relevant components of equation (19) on the basis of empirical data. In my 1974 paper, estimates of the overall elasticity of "all felony offenses" with respect to probability and severity of imprisonment for these felonies, derived through a cross-state regression analysis using 1960 data for the United States, show the latter to be about unity. Even exaggerated estimates of the actual incapacitative effect of imprisonment, based on the assumption that $p = 1/3$ (i.e., one of every three offenders at large is apprehended and imprisoned every year), and, say, $\alpha = \beta$, show that incapacitation cannot explain even 25 percent of the observed elasticity. More realistic estimates would place the latter proportion for most crimes well under 10 percent (see the Appendix). It appears therefore that in practice the overwhelming portion of the total preventive effect of imprisonment is attributable to its pure deterrent effect.

V. On the Choice of Optimal Sanctions

The optimal deployment of alternative means of crime control cannot be based merely on their relative efficacy in reducing offenses, but must involve consideration of their relative costs. The analysis here will focus on the optimal choice among alternative sanctions, using the model of optimal public enforcement outlined in Section I.

Fines and related monetary exchanges exert a purely deterrent effect. In contrast, imprisonment, detention, and probation render both incapacitative and deterrent services. By equation (19) the latter sanctions

[19] See, for example, Shinaar and Shinaar; David Greenberg.

must exert a total preventive effect that is either equal to or greater than the purely deterrent effect of a fine of equal cost to the offender.

Imprisonment and fines are associated, however, with different social costs as well—a point which is central to Becker's important analysis of the case for fines. Whenever feasible, an optimal fine would amount to a transfer payment made by the offender to compensate the rest of society for the external costs inflicted through his criminal conduct. The net resource costs to society from fines are then the costs of the "collection agency." In contrast, imprisonment is a nontransferrable, noncompensating payment made by the offender in the form of foregone earnings and other restrictions on personal freedoms during the period of incarceration. The part of an offender's foregone income that is derived from legitimate activities is, of course, a genuine social cost as well. In addition, the administration and maintenance of a prison system involves considerable expenditures of resources. In equation (1) the cost to society imparted by any sanction imposed on the offender is formally captured by the multiplier b. As long as fines are feasible $b(m)>b(f)\geqslant 0$, where m and f stand for the monetary equivalents of imprisonment and fines, respectively.

If imprisonment and fines were constrained to be mutually exclusive, the choice of the optimal value of either sanction would be determined through minimization of equation (1), generally expressed

$$L(p,t)=\Delta(q)+C(q,p)+b(t)ptq, \qquad t=m,f$$

with respect to m and f separately. For any given value of p, the optimality conditions relating to f and m are given by

$$(20) \qquad \Delta_q+C_q=b(m)pm(E_m-1)$$

and

$$(21) \qquad \Delta_q+C_q=b(f)pf(E_f-1)$$

where $E_t\equiv 1/\varepsilon_t$, and $\varepsilon_t\equiv -\partial\, ln\, q^*/\partial\, ln\, t$ denotes the elasticity of the *equilibrium* crime rate with respect to t, $t=m,f$. The determinants of ε_m and ε_f can be inferred from equation (19): While ε_m would include both terms on the right-hand side of equation (19), ε_f would be represented by the second term alone. Note that in this specification of the social loss function, where t is not an argument in the cost function of direct law enforcement activity, $C(q,p)$, and where no distributive effects of enforcement are considered as part of the social target function, equations (20) and (21) can be satisfied only if both ε_m and ε_f are less than unity. This restriction does not apply, however, to the magnitude of the elasticities of the supply-of-offenses function $S(\pi)$ (α in equation (7)), which by equation (19) is free to vary between zero and infinity. Put differently, the restriction $\varepsilon_m<1$ does not imply that in equilibrium *offenders* cannot be highly responsive to incentives, as previous analyses seem to suggest.

Equations (20) and (21) imply that, when forced to invoke either f or m, the law enforcement authority would make its choice according to whether

$$(22) \qquad \frac{(E_m-1)}{(E_f-1)} \begin{matrix} > \\ = \\ < \end{matrix} \frac{b(f)f}{b(m)m}$$

This result can be interpreted as follows: for any target level of offenses q^0, given the value of p, a fine would dominate imprisonment as an optimal sanction, provided that its potentially lower overall preventive effect is more than offset by its relatively lower social cost. More generally, if imprisonment and fines could be imposed jointly, their values would be chosen so as to minimize

$$(23) \qquad L(p,f,m)=\Delta(q)+C(q,p) + \left[b(m)m+b(f)f\right]pq$$

and the optimal combination of f and m, given p, would be required to satisfy

$$(24) \qquad \frac{m}{f}=\frac{b(f)\varepsilon_m}{b(m)\varepsilon_f}$$

Both equations (22) and (24) point to the superiority of fines in those cases where fines do not exhaust an offender's financial constraint, and $b(f) \approx 0$. The superiority of fines is further apparent in the particular case where the elasticity of the market demand-for-offenses schedule (β in equation (8)) is nil. As equation (19) indicates, the elasticity of the equilibrium crime rate with respect to *any* means of crime control would be zero in this case. The superiority of monetary fines would then be unqualified because only compensation to victims could internalize the external costs of crime.

These specific illustrations underscore the dual role of monetary fines, both as a means of crime prevention and as a Pigouvian tax. The general analysis at the same time modifies Becker's assertion that maximization of social welfare requires the exclusive use of fines whenever they are feasible. Since, in general, $b(f) > 0$, the analysis shows that even when feasible, fines should be replaced by, or used in conjunction with, an incapacitating penalty if ε_m were sufficiently greater than ε_f. By equation (19), the difference between the two is proportional to $[pT/(1+pT)] \cdot [\beta/(\alpha+\beta)]$.

VI. Some Illustrations

A. *Discriminating Penalties*

Since public law enforcement is carried out under state monopoly, it would be optimal for law enforcement authorities to impose different penalties on different groups of offenders if the marginal social return from enforcement differed systematically across these groups. Equation (20) indicates, for example, that the optimal severity of imprisonment would be higher, the higher is ε_m. Becker has argued on the basis of this condition that insane and young offenders, or perpetrators of unpremeditated crimes whose responsiveness to incentives is presumed to be low, should receive lighter penalties than other, more responsive offenders. The analysis of Section V changes this conclusion, because the magnitude of ε_{mj} associated with different groups of offenders j is determined by three distinct effects: a) a deterrent effect; b) an incapacitative effect; and c) the interplay of supply and demand forces. While the pure deterrent effect of imprisonment is an increasing function of a group's responsiveness to incentives α_j, its pure incapacitative effect under a segmented market structure is a decreasing function of α_j. It is thus possible, at least in principle, that ε_{mj} and α_j would not be positively correlated.

The conclusion that insane, nonpremeditating, and "hardened" offenders should be given relatively lighter penalties holds unqualifiably only under the constraint that all offenders are to be punished through purely deterring sanctions. An optimal policy would not exempt unresponsive offenders from punishment, but punish them through incapacitative penalties. Moreover, it might even pay to punish them relatively *more* severely. Little responsiveness to incentives, then, is no justification for little punishment, but rather for punishment of a different kind.

B. *The Control of Crimes Against Persons*

It is frequently asserted, although not systematically documented, that perpetrators of crimes of passion are less responsive to incentives than other offenders. While undoubtedly valid in particular cases, this assertion need not hold in general.[20] It is, however, possible that the distribution of individual preferences for such crimes is subject to marked discontinuities which can contribute to an inelastic shape of the aggregate supply-of-offenses schedule about typical equilibria positions. In addition, it is possible that the market derived-demand schedule for such crimes is quite elastic. By these considerations the efficacy of rehabilitation and incapacitation may indeed be higher in connection with crimes against persons.

Note, however, that a prerequisite for any method of individual control to be efficacious at both the individual and the aggregate levels is that there be a positive and

[20] Samuel Yochelson and Stanton Samenow report evidence, based on psychological observations during treatment, that a majority of crimes of passion are in fact nonspontaneous and deterrable. Also see the analyses and evidence reported in my 1975 and 1977 papers and in Kenneth Wolpin.

significant probability of individual *recidivism*. Many crimes against persons, especially murders and assaults, are committed as a result of personal frictions under unique personal circumstances that have low probabilities of recurrence once the crime is committed. Methods of individual control would then inherently be productive only in connection with those perpetrators of crimes against persons whose probability of recidivism is high.

C. *Victimless Crimes*

The supply elasticities α_i of "victimless crimes," such as gambling, loan sharking, prostitution, and the sale of all illicit goods, are likely to be particularly high as these criminal enterprises share many of the characteristics of business endeavors in legitimate markets. No one would suggest that the act of shutting off a gasoline station because of violation of safety or health codes, or its conversion to a bicycle shop, can per se result in a comparable reduction in the aggregate amount of gasoline sold in the relevant local market. The reduced supply by the obstructed station will almost surely be replaced by increased production by competitors and jobbers. The same goes for prostitution and transactions in drugs. And because the consumers patronizing these businesses may have relatively inelastic demands, any law enforcement crackdown on these businesses would mainly hike the prices of the commodities involved without affecting markedly the volume of transactions. Monetary fines or taxes would produce both the maximum amount of crime prevention via deterrence, and compensation for members of society who, in various personal ways, may be victimized by the activities in question.

D. *A Parallel between Rehabilitation and Retraining Programs*

The equilibrium analysis developed in this paper is applicable in explaining not only the evidence concerning the efficacy of rehabilitation programs for offenders, but also the evidence emerging from evaluation studies of retraining programs of adult workers in specific legitimate industries (see, for example, Charles Perry et al., pp. 183–200). Public retraining of workers for superior jobs or skills in specific industries subject to high degrees of technological innovations amount to an attempt to reshape the shifting supply schedules of workers to these jobs. The latter reflect the minimum wage differentials required by individual workers to enter (or reenter) the submarkets for the relatively higher skills, in view of the additional investments necessary. If subsidized retraining is successful in imparting the required knowledge, it will enable the retrained workers to compete with newly trained workers for the skilled positions available at the going market wages. However, since the retraining programs do not affect the industries' derived demand schedules for the specific skills involved, the total employment of these skills (hence the actual integration of retrained workers) would not be markedly affected if the supply schedules of the specific skills were sufficiently elastic, and those of the derived demand schedules were low.

VII. A Concluding Remark

I do not mean to suggest that methods of individual control should be abolished; rehabilitation may serve a variety of social objectives, not all of which include crime prevention. Incapacitation would be necessary for specific types of offenses or offenders where the extent of individual responsiveness to incentives is low and the rate of recidivism is high. My analysis shows, however, that, in a large class of cases, efficient crime control requires only the imposition of deterring punishments or the promotion of general legitimate earning opportunities, without any attempt at individual control.

Appendix: A General Analysis of Maximum Removal Effects

The following analysis is based on the original model of the incapacitative effect of imprisonment developed in my 1974 paper. Let the total population in a given community be represented by N, with N growing over time at the geometrical rate g. Given an equilibrium rate of criminal returns $\pi^*=\pi_0$ and other determinants of participation in criminal activity, a fraction $s^e(\pi_0)$ of the total population would be attracted to the market for offenses in any given period. The stock of offenders in t is thus given by

$$\text{(A1)}\qquad S_t^e(\pi_0)=s^e(\pi_0)N_t$$
$$=s^e(\pi_0)N_0(1+g)^t$$

With individual methods of crime control effectively used, the stock of offenders at large is given by

$$\text{(A2)}\qquad \theta_t=S_t^e(\pi_0)-R_t$$

where R_t denotes the number of offenders actually removed.

Assume that a fraction $p<1$ of θ_t is apprehended and effectively removed each period for a duration of T periods. Then, by application of the analysis in my 1974 paper, the steady-state effective stock of offenders at large (per capita) can be easily derived:

$$\text{(A3)}\qquad \bar{\theta}=\frac{1}{1+p\sum\limits_{\tau=1}^{T}(1+g)^{-\tau}}s^e(\pi_0)$$

If an average offender at large commits $k(\pi_0)$ offenses per period when $\pi=\pi_0$, the (per capita) supply-of-offenses function can now be specified as

$$\text{(A4)}\qquad q=\frac{k(\pi_0)}{1+p\sum\limits_{\tau=1}^{T}(1+g)^{-\tau}}s^e(\pi_0)$$
$$=A_0(pT)s(\pi_0)$$

where $$A_0=\frac{1}{1+p\sum\limits_{\tau=1}^{T}(1+g)^{-\tau}}$$

and $s(\pi_0)\equiv k(\pi_0)\cdot s^e(\pi_0)$. Clearly, equation (A4) is of the format of the supply-of-offenses function analyzed in the text, with $s(\pi_0)=\pi^\alpha$, and with A_0 reducing to $1/(1+pT)$ if $g=0$. The effect of a percentage increase in the fraction of offenders removed from the market in a given period on the steady-state value of q will therefore be given by

$$\text{(A5)}\qquad \sigma^{max}\equiv-\left.\frac{\partial\, ln\, q}{\partial\, ln\, p}\right|_{\pi^*=\pi_0}$$
$$=\frac{p\sum\limits_{\tau=1}^{T}(1+g)^{-\tau}}{1+p\sum\limits_{\tau=1}^{T}(1+g)^{-\tau}}=\left.\frac{\partial\, ln\, q}{\partial\, ln\, T}\right|_{\pi^*=\pi_0}$$

which reduces to $\sigma^{max}=pT/(1+pT)$ if $g=0$ (see equation (17)). By substituting rp and L, as defined in section III, for p and T in equation (A5) and setting $g\approx0$, equation (14) is also immediately derived. Note that the assumption $g=0$ overstates the value of σ^{max} whenever $g>0$.

Illustration: According to *Characteristics of State Prisoners, 1960,* the average length of time spent in state prisons by offenders upon their first release from prison in 1960 for all index crimes is estimated at 30.75 months or 2.56 years.[21] A measure of the probability that an offender at large be apprehended and imprisoned for these crimes in 1960, calculated as the ratio of offenders committed to state prisons C^0 to the total number of offenses known to police Q^0, sets p at 0.028.[22] The values of σ^{max} based on these values of p and T is 0.066. In contrast, the estimated elasticities of the same offenses with respect to p and T, based on a 1960 cross-state regression analysis, are found to average about unity in absolute magnitude (see my 1974 paper, Table 5). Clearly, an estimate of

[21] This is the weighted average of the actual times served for the specific index-crime categories, weighted by the data on releases. The index crimes include murder, rape, aggravated assault, robbery, burglary, larceny, and auto theft.

[22] The value of C^0 is calculated from *Characteristics of State Prisoners, 1960,* Table A1, and Q^0 from *Uniform Crime Reports* (*UCR*), Table 2.

TABLE 1—ESTIMATES OF THE INCAPACITATIVE AND DETERRENT EFFECTS OF IMPRISONMENT BASED ON 1960 DATA

	p	T	$\sigma^{max}=\frac{pT}{1+pT}$	$\theta=\frac{\beta}{\alpha+\beta}\sigma^{max}$	ε_p	$s=1-\theta/\varepsilon_p$
Category	(1)	(2)	(3)	(4)	(5)	(6)
All Offenses	.028	2.56	.066	.033	.991	.967
	.10[a]	2.56	.20	.10	.991	.90[a]
	.20[a]	2.56	.33	.17	.991	.83[a]
	.33[a]	2.56	.46	.23	.991	.77[a]
Specific Crimes:						
Murder	.398	10.12	.801	.400	.852	.531
Rape	.227	3.73	.458	.229	.896	.744
Aggravated Assault	.030	2.08	.059	.029	.724	.960
Robbery	.084	3.53	.229	.114	1.303	.913
Burglary	.024	2.05	.047	.023	.724	.968
Larceny	.022	1.65	.035	.017	.371	.954
Auto Theft	.021	1.78	.036	.018	.407	.936

Sources: Data for columns (1) and (2) are given in fnn. 21 and 22. In column (4), I set $\alpha=\beta$. Column (5) lists empirical estimates of the elasticity of specified offenses with respect to the probability of imprisonment ε_p as reported in my 1974 paper, Tables 4 and 5. Column (6) represents estimates of the share of deterrence in ε_p.

[a]Hypothetical estimates based on arbitrary values of p.

p based on C^0/Q^0 may be seriously biased in both an upward and a downward direction. The desired measure of the probability p may be approximated by C^0/θ, where θ is the number of offenders at large, or $K/Q=kC^0/K\theta$, where k is the average number of offenses committed by an offender at large (see my 1974 paper, p. 124). Clearly, while $C^0<K$, if $k>1, Q>Q^0$ because the number of offenses reported is substantially lower than the true number of offenses committed.[23]

Alternative estimates of σ^{max} and the actual incapacitative effect of imprisonment can be obtained (see Table 1) by placing arbitrary values on the magnitude of p. If one were willing to assume that as much as one in ten offenders at large is actually apprehended and imprisoned in a given year, σ^{max} will be 0.20 and the actual incapacitation effect $\sigma^{max}(\beta/\alpha+\beta)$, with $\alpha=\beta$, would be 0.10, or just about 10 percent of the actually estimated elasticity. It is clear that any reasonable estimate of p implies that the bulk of the empirically estimated effect of imprisonment on crime (represented in Table 1 by ε_p) is due to deterrence, especially if α were high.

Note that the share of deterrence in ε_p as estimated by s in Table 1 may be understated both because σ^{max} is estimated under an assumed zero population growth, and because no attempt is made to deduct from the calculated incapacitative effect of imprisonment its hardening effect on released offenders. Also note that the estimate of s may be understated especially in the case of murder because my measure of the relevant p for murder is biased upward: it is calculated on the assumption that any potential murderer at large commits one murder every year both before and after his imprisonment (see fn. 23 and Section VI, Part B).

[23]FBI data from 1960, which provide estimates of the probability of arrest relying on both an offense-based measure (percentage of offenses cleared by arrest) and an offender-based measure (persons charged relative to offenses known) show the latter estimate to be 24 percent lower than the former. (See *UCR*, Table 9.) In contrast, some estimates of the extent of underreported crime show reported index crimes to be 50 to 75 percent lower than the "true" number of crimes in 1964. (See President's Commission on Law Enforcement and Administration of Justice, pp. 17, 18.) Note that in the case of murder k may be substantially less than unity for "offenders at large" (see Section VI, Part B.). Hence my estimate of p for murder might be seriously overstated.

REFERENCES

W. C. Bailey, "Correctional Outcome: An Evaluation of 100 Reports," in Leon Radzinowicz and Marvin E. Wolfgang, eds., *Crime and Justice*, Vol. 3, New York: Basic Books 1971.

G. S. Becker "Crime and Punishment: An Economic Approach," in his and William M. Landes, eds., *Essays in the Economics of Crime and Punishment*, New York: Columbia University Press 1974.

P. J. Cook, "The Correctional Carrot: Better Jobs for Parolees," *Policy Analysis*, Winter 1975, *1*, 12–51.

J. M. Chaiken, M. W. Lawless, and K. A. Stevenson, "The Impact of Police Activity on Crime: Robberies on the New York City Subway System," rept. no. R-1924, Rand Institute, New York 1974.

I. Ehrlich, "Participation in Illegitimate Activities: An Economic Analysis," in Gary S. Becker and William M. Landes, eds., *Essays in the Economics of Crime and Punishment*, New York: Columbia University Press 1974.

______, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *Amer. Econ. Rev.*, June 1975, *65*, 397–415.

______, "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," *J. Polit. Econ.*, Aug. 1977, *85*, 741–88.

______, "The Economic Approach to Crime: A Preliminary Assessment," in Sheldon L. Messinger and Egon M. Bittner, eds. *Criminology Review Yearbook No. 1*, Beverly Hills: Sage Publications 1979.

______ **and G. S. Becker,** "Market-Insurance, Self-Insurance, and Self-Protection", *J. Polit. Econ.*, July/Aug. 1972, *80*, 623–48.

______ **and J. Gibbons,** "On the Measurement of the Deterrent Effect of Capital Punishment and the Theory of Deterrence," *J. Legal Stud.*, Jan. 1977, *6*, 35–50.

D. Greenberg, "The Incapacitative Effect of Imprisonment: Some Estimates," *Law Soc. Rev.*, Summer 1975, *9*, 580–81.

R. G. Hood, "Research on the Effectiveness of Punishments and Treatments," in Leon Radzinowicz and Marvin E. Wolfgang, eds., *Crime and Justice*, Vol. 3, New York: Basic Books 1971, 159–82.

Douglas Lipton, Robert Martinson, and Judith Wilks, *The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies*, New York: Praeger 1975.

C. R. Perry et al., "The Impact of Government Manpower Programs," Manpower and Human Resources Studies, No. 4, Univ. Pennsylvania 1975.

Leon Radzinowicz, *Ideology and Crime*, New York: Columbia University Press 1966.

J. Robison, and G. Smith, "The Effectiveness of Correctional Programs," *Crime Delinquency*, Jan. 1971, *17*, 67–80.

S. Shinaar and R. Shinaar, "A Simplified Model for Estimating the Effects of the Criminal Justice System on the Control of Crime," unpublished paper, Sch. Engineering, City College, New York 1974.

W. Vandaele, "The Economics of Crime: An Econometric Investigation of Auto Theft in the United States," unpublished doctoral dissertation, Univ. Chicago 1975.

Hal R. Varian, *Microeconomic Analysis*, New York: Norton 1978.

Leslie T. Wilkins, *Evaluation of Penal Measures*, New York: Random House 1963.

James Q. Wilson, *Thinking About Crime*, New York: Basic Books 1975.

K. Wolpin, "Capital Punishment and Homicide in England: A Summary of Results," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 422–27.

Samuel Yochelson and Stanton E. Samenow, *The Criminal Personality, Vol. I: A Profile for Change*, New York: Jason Aronson 1976.

President's Commission on Law Enforcement and Administration of Justice (*PCL*), *Crime and its Impact—An Assessment*, Task Force Report, Washington, D.C.: U.S. Government Printing Office 1967.

U.S. Department of Justice, *Characteristics of State Prisoners, 1960*, National Prisoner Statistics, Washington, D.C.: U.S. Government Printing Office 1960.

U.S. Department of Justice, Federal Bureau of Investigation, *Uniform Crime Reports*, Washington, D.C.: U.S. Government Printing Office 1961.

The Future Price of Houses, Mortgage Market Conditions, and the Returns to Homeownership

By SUSAN I. RANNEY*

The purpose of this paper is to provide a simple framework in which to analyze the impact of a perfectly anticipated increase in the future selling price of houses on current house purchase decisions in the presence of capital market imperfections. Since, implicitly, it is the after-tax selling price of houses that is considered, this analysis is equivalent to one of a decrease in the capital gains tax on houses at retirement. In this paper assumptions about homeownership, mortgage markets, and liquidity constraints are added to a life cycle model of asset accumulation, taking the date of house purchase as exogenous. It is shown that the impact of an exogenous increase in the expected future price of houses depends on current mortgage market conditions, the future price of houses relative to the current price, and the ratio of liquid assets to future labor income, as well as preferences. Also, a simple model of a market for houses is presented to analyze the impact of an increase in the expected future price of houses on the current market-clearing price, the mortgage market, and the redistribution of the housing stock to consumers according to their wealth and liquidity characteristics.

In the life cycle model as exposited by Franco Modigliani and Robert Brumberg, an increase in consumption in one period requires a decrease in consumption in another period. When owner-occupied housing is added to the model the tradeoff between consumption in different periods becomes more complex. Purchasing a larger house entails an increase in the consumption of housing services and an increase in assets held in the form of a house, but a decrease in either nonhousing consumption or the holdings of an alternative asset, or both. The response of the consumer to price or income changes is also limited by financial market constraints. Capital market imperfections have been introduced into the life cycle model by a number of authors (see, for example, Lester Thurow, Thayer Watkins, Thomas Russell, Clark Wiseman, and Christopher Pissarides). And, several authors have explicitly considered the impact of capital market imperfections on the house purchase decision (see William Dolde and James Tobin, William Poole, Donald Lessard and Modigliani, and Dolde). Most similar to this study, although developed independently, Roland Artle and P. Varaiya's paper includes the theoretical incorporation of the house-purchase decision into a life cycle model. They consider the problem of choosing the date of house purchase while treating the level of consumption of housing services as exogenous. In the life cycle model in this paper, the consumer chooses the size of house to purchase at an exogenous date.

To use the life cycle model in the framework of a market of houses, it is assumed that the stock of houses and the prices of all other goods, including the rate of return on the alternative asset, are fixed. All consumers are identical, except perhaps for initial wealth. As noted by Irving Fisher (p. 325), in a world of perfect foresight and perfect markets the rates of return on all assets are equated in equilibrium. If markets are perfect, with the stock of houses and other prices fixed, the present price of houses will rise by an amount equal to the present value of an exogenous increase in a future price of houses. When mortgage market imperfections are introduced, this relationship no longer holds, and a future price increase may result in a redistribution of the current stock

*Assistant professor of economics, University of Michigan. This paper is an extension of a chapter of my doctoral thesis at the University of Wisconsin-Madison. The helpful comments of Donald D. Hester, Charles A. Wilson, Hal Varian, and George Borts are gratefully acknowledged.

among consumers with differing wealth and liquidity characteristics.

I. A Model of the House-Purchase Decision

In the analysis it is assumed that the lifetime utility of a household is a function of the consumption of housing services and the consumption of other goods and services in every instant of time between the initial time and retirement; in addition, lifetime utility is assumed to be a function of the value of assets at the date of retirement. Retirement assets are equal to the value of the house plus the stock of savings at retirement. Savings are defined as financial assets, exclusive of house equity and mortgage debt, and are assumed to earn an exogenous rate of return that is constant through time. The consumer is free to choose the size of house to be purchased at time zero, and can choose the fraction of house that is to be financed by mortgage borrowing up to an institutionally imposed maximum. The mortgage interest rate is assumed to be greater than the rate of return of savings.

There are two additional constraints on the path of consumption over the lifetime. First, the stock of accumulated savings at any point in time can never fall below zero. Second, once a house is bought, the transactions costs of selling it are assumed to be such that it is not sold until retirement. In this model, the physical characteristics of a house are assumed to be fixed over time. The date of purchase, 0, and the date of retirement, T, are assumed to be exogenous, and the house is always sold at retirement. The consumer is acting under perfect certainty.

A. *The Basic Model*

In the model, U is an instantaneous utility function of consumption in time t. The function U is assumed to be increasing, twice differentiable, and concave in its arguments. It is assumed that the marginal utility of each good consumed approaches zero, holding the quantity of the other good constant.

The function F measures utility derived from the value of assets at retirement; it is assumed to be increasing, twice differentiable, and concave. Notation for variables and parameters is listed in Table 1.

TABLE 1

Exogenous Variables and Parameters

I_t = labor income, time t

γ = constant which converts house size into a flow of housing services

r = interest rate paid on savings

r_m = mortgage interest rate

$\hat{r}_m$ = constant which converts the total mortgage into a constant, continuous mortgage payment:

$$\hat{r}_m=\frac{r_m}{\left(1-e^{-r_m T}\right)}$$

β = maximum fraction of value of the house that can be mortgaged at time 0

P_t = price of other goods and services, time t

P_H = price of house units, time 0

P_{HT} = price of house units, time T

Endogenous Variables

A_t = consumption of housing services, time t

B_t = consumption of other goods and services, time t

S_t = stock of savings, time t, exclusive of equity in the house

DS_t = derivative of S_t with respect to time

S^+ = the stock of savings the instant after the house purchase

S = the initial stock of savings or wealth, before the house purchase

H = size of house purchased

b = fraction of house that is mortgaged at time 0

W = wealth at retirement

The problem facing the consumer is

$$(1)\qquad \max\int_0^T U(A_t,B_t)\,dt+F(W)$$

subject to

(a) $S^+=S-(1-b)P_H H,\ S^+>0$

(b) $0\leqslant b\leqslant\beta$

(c) $A_t=\gamma H$

(d) $DS_t=I_t+rS_t-b\hat{r}_m P_H H-P_t B_t$

(e) $S_t\geqslant 0$ for all t

(f) $W=S_T+P_{HT}H$

The consumer maximizes utility over the lifetime, choosing the house size and size of mortgage at time 0. The mortgage is paid off at a constant rate from 0 to T, and the remaining income is either saved or spent on other goods and services.[1] The constraints

[1] Both the return on the alternative asset and the mortgage payments should be viewed as after-tax values. For a more complete treatment of taxes in a similar model see my dissertation.

can be summarized as follows:

(a) Immediately after the house purchase, the stock of savings must be nonnegative and is equal to the initial stock of savings minus the downpayment on the house.

(b) The fraction of the house mortgaged cannot be chosen less than zero or greater than the institutionally imposed maximum.

(c) After the house purchase, the flow of housing services at any instant of time is equal to an exogenously given fraction of the size of the house purchased.

(d) After the house purchase, saving is equal to total income (labor plus interest) minus expenditures on mortgage payments and other goods.

(e) There is a nonnegativity constraint on financial savings up until retirement. A binding liquidity constraint at T is discussed in Section II, Part C.

(f) Real wealth at retirement is equal to the value of the house plus the final stock of nonhouse savings. A price index for retirement wealth that depends on the price of houses at T is considered in Section II, Part D.

B. *Reduction of the Model to a Static Model with Two Goods*

By making one additional assumption, the model can be reduced to a static-maximization problem with two goods:

$$(1') \qquad S_t > 0 \text{ for all } t, \quad 0 < t < T$$

Assumption (1′) states that, in the solution to the maximization problem, the household chooses to allocate expenditures such that the nonnegative savings constraint is not binding, except perhaps at the date of purchase. When the solution satisfies (1′), the consumer is allocating expenditure optimally over the interval $(0, T)$. When (1′) does not hold, the consumer might prefer to borrow against future income in order to consume more in the earlier part of a time period, but is unable to do so.

With (1′) the choice of the consumption stream within the interval $(0, T)$ does not depend on the distribution of labor income within the interval, but only on the present value of the stream of income. Therefore, the only binding budget constraint in $(0, T)$ is the aggregated intertemporal budget constraint:

$$(2) \qquad S_T e^{-rT} = S^+ + \int_0^T (I_t - P_t B_t - \hat{r}_m b P_H H) e^{-rt}\, dt$$

Or, letting

$$Y = \int_0^T I_t e^{-rt}\, dt \quad \text{and} \quad k = \int_0^T \hat{r}_m e^{-rt}\, dt$$

then

$$(3) \qquad S_T e^{-rT} = S^+ + Y - kbP_H H - \int_0^T P_t B_t e^{-rt}\, dt$$

The present value at 0 of labor income earned after the house purchase is equal to Y. The variable k is equal to the present value of the mortgage payments for one dollar of mortgaged house. Equation (3) states the the present value of the final stock of savings is equal to the stock of savings immediately after the purchase plus the present value of total saving from 0 to T.

Next, define

$$(4) \quad V(H, Q) = \max_{B_t, W} \int_0^T U(\gamma H, B_t)\, dt + F(W)$$

$$\text{subject to} \qquad Q = \int_0^T P_t B_t e^{-rt}\, dt + We^{-rt}$$

At the maximum, Q is equal to the present value of expenditure of nonhousing goods and services from 0 to T plus total wealth at retirement, or "other goods." $V(H, Q)$ is total lifetime utility written as a concave function of H and Q, where P_t from 0 to T and r enters as exogenous variables.[2] Once H and Q are determined, the first-order conditions for a maximum in (4) determine the allocation of funds to W and to B_t over time.

[2] Properties of $V(H, Q)$ are discussed in my dissertation .

To determine H and Q, (3) and (4) are substituted into (1), using (1′). The lifetime utility-maximization problem can be stated:

$$(5) \qquad \max V(H,Q)$$

subject to
(a) $S+Y+PH=(1-b+bk)P_H H+Q$
(b) $S-(1-b)P_H H \geqslant 0$
(c) $0 \leqslant b \leqslant \beta$
where $P=P_{HT}e^{-rt}$.

Constraint (a) states that lifetime wealth, consisting of initial wealth, labor income, and the selling value of the house, is allocated to expenditure on the house and expenditure on other goods. Constraint (b) is the nonnegativity constraint for the stock of savings immediately after the house purchase, and (c) is the down-payment constraint.

The first-order conditions can be expressed as

$$(6) \quad (V_H+PV_Q)/P_H=kbV_Q+\lambda(1-b)$$

$$(7) \qquad \begin{array}{lll} \lambda=kV_Q & \text{if} & 0<b<\beta \\ \geqslant & & b=\beta \\ \leqslant & & b=0 \end{array}$$

$$(8) \qquad \lambda \geqslant V_Q \text{ with equality if } S-(1-b)P_H H>0$$

where $V_i=\partial V/\partial i$ for $i=H$, Q, and $\lambda=$ marginal utility of the initial stock of savings.

Condition (6) equates the marginal utility of expenditure on the house with the opportunity cost. The marginal utility of expenditure is composed of both the value of the marginal utility of additional housing services, and the marginal utility of expenditure of the revenue from selling the house. The opportunity cost of expenditure can also be broken down into two parts: that corresponding to the portion paid in downpayment and the portion paid in mortgage payments. Condition (7) determines the proportion of the house that is mortgaged. If b is unconstrained then it is chosen so as to equate the opportunity cost of a dollar of downpayment with the opportunity costs of a dollar of mortgage. Condition (8) states that the shadow prices of initial savings and postpurchase income are equated if it is optimal to carry financial assets from the pre- to the postpurchase period.

C. *Diagrammatic Analysis*

To examine the model diagrammatically, first the budget set for the maximization problem is described. Note that the consumer never enters the mortgage market unless the entire stock of savings at the date of purchase is put into the house. This follows from three assumptions: (a) the mortgage rate is greater than the rate of return on savings; (b) there is no utility derived directly from holding savings except at time T; and (c) with (1′), there are no restrictions on savings arising form the shape of the income stream between the date of purchase and T.

Thus there are three possible constraints that may restrict the household's allocative decisions. These determine the feasible budget set. First, if the entire house is financed by initial savings, the budget constraint can be written as

$$(9) \qquad Y+S-(P_H-P)H-Q \geqslant 0$$

This inequality is depicted by lines a in Figure 1.

If the mortgage market is used the budget constraint becomes

$$(10) \quad Y+kS-k(P_H-P)H-Q \geqslant 0$$

This is shown by lines b in Figure 1.

Last, the minimum down-payment constraint

$$(11) \qquad S-(1-\beta)P_H H \geqslant 0$$

is depicted by lines c in Figure 1. The feasible sets are indicated by the shaded areas.

In Figure 1A it is assumed that $P_H>P$, implying the effective price of housing services is always positive. Given this budget set, the possible points of maximization can be characterized by four cases, depending on the preference of the consumer. In case I, the down-payment constraint is binding, so that the stock of savings immediately before the purchase is equal to a fixed proportion of the house. In case II, the down-payment con-

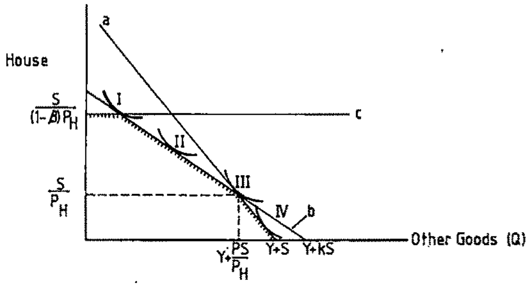


FIGURE 1A

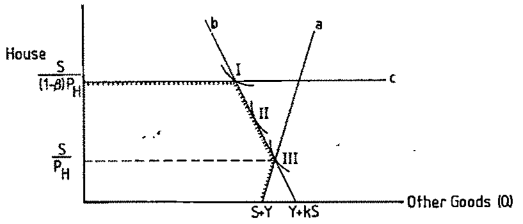


FIGURE 1B

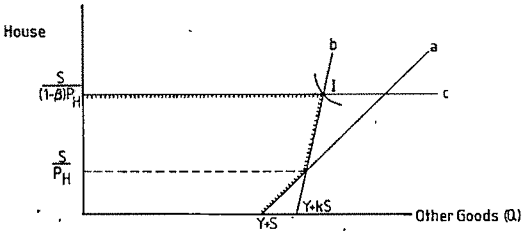


FIGURE 1C

straint is not binding, but the consumer does hold a mortgage. Here, the proportion of the house mortgaged is not established by the constraint. In case III, the house is no longer mortgaged, but all of initial wealth is spent on the house. Thus expenditure on other goods is equal to the revenue from the house sale plus labor income. In case IV, none of the financial market imperfections has any impact. The mortgage market is not used and savings immediately after the house purchase are positive, so that only the lifetime budget constraint is binding.

The budget set for $P > P_H$ is shown in Figures 1B and C. When $P > P_H$, the price of housing services is negative as long as the mortgage market is not used. The consumer therefore always increases house size until the nonnegativity constraint on the stock of savings immediately after the purchase becomes binding, and no consumer ever chooses to be in case IV. The house yields a higher rate of return in appreciation alone than that earned on financial savings and, in addition, provides housing services.

If $kP_H > P > P_h$, as in Figure 1C, the price of housing services is negative, even with a mortgage, and the consumer increases house size until the down-payment constraint becomes binding. Only the point corresponding to case I on the budget set will ever by chosen.

In the analysis that follows in this section it is assumed that $P < P_H$. In this way the most general budget set is used, allowing for the existence of all four cases. It is also assumed that H and Q are normal goods, when $V(H, Q)$ is maximized subject only to a linear budget constraint.

With these assumptions, an expansion path in (H, Q) space for an increase in S can be drawn. In Figure 2 an expansion path designated by d is drawn for the case of homothetic preferences. The two kinks in the budget constraint shown in Figure 1 move out along two lines in Figure 2 as the budget set shifts out with an increase in S. One kink is the result of the change in the price of H relative to Q when the consumer switches from paying for the house out of savings to financing the house through a mortgage. Since the mortgage interest rate is greater than the rate of return on savings, the consumer begins mortgaging only when total initial wealth is used in the down payment. So the kink always occurs where S is equal to $P_H H$; or $Q = PH + Y$, as shown by line e in Figure 2.

The second kink results from the down-payment constraint. The amount of post-purchase income that can be allocated to the house through mortgaging is limited by the down-payment constraint. When the down-payment constraint is binding, there is a lower bound on the amount of Q consumed, so that $Q \geqslant Y + (P - P_H \beta k)H$. This is line f in Figure 2. As S increases, the kink moves out along this line. It is drawn for the case where $P < P_H \beta k$, but since βk could be

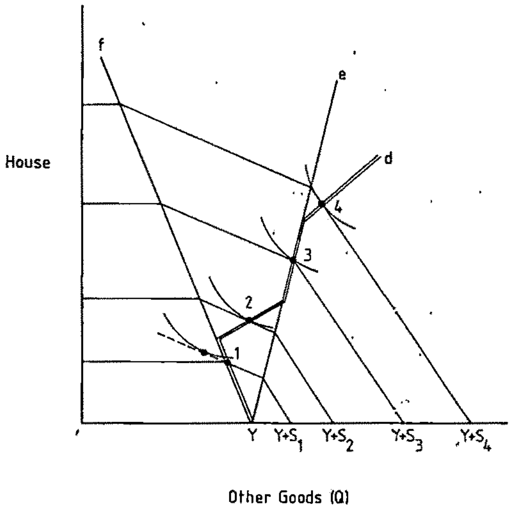


FIGURE 2.

greater or less than one, (f) could be upward sloping even with $P_H > P$.

In Figure 2, points 1, 2, 3, and 4 correspond to households in cases I, II, III, and IV, respectively. The case I household is limited by the down-payment constraint. As initial savings increase, the household moves along f, increasing house size. SInce it costs more in terms of future goods consumption to mortgage than to pay down, as S increases the household may eventually find that it is no longer optimal to mortgage the maximum proportion of the house, and move into case II. But, in case II, as S increases the percentage of the house mortgaged decreases, and the consumer may reach case III, where b is equal to zero.

In case III, the price of other goods relative to the initial price of houses is too high for the consumer to retain any of the initial wealth in savings after the purchase. But the cost of a mortgaged house relative to the price of other goods is too high for the consumer to borrow. Starting at S_3, an increase in S results in an increase in both H and Q, with a smaller increase going to Q. As S continues to increase, the consumer may eventually find that it is optimal to put some of the initial wealth into financial savings and enter case IV. As the stock of savings continues to increase from S_4, the choice of H and Q is not limited by financial market constraints.

II. An Increase in the Future Selling Price of Houses

In this section, the partial equilibrium impact of an increase in the price of houses at retirement is analyzed, and a simple market model is used to examine the impact of an increase in the future selling price on the initial buying price. In the market model, it is assumed that the supply of houses is fixed but infinitely divisible at time 0, and that P_H adjusts to equate demand with supply. The supply of all other goods is assumed to be infinitely elastic so that all other prices are invariant to changes in the price of houses.

A. *Analysis of the Four Cases*

The impact of an increase in P on the budget set is shown in Figure 3. Here it is assumed that the consumer is in the same case before and after the price change. In case IV, the mortgage market is not used and savings immediately after the house purchase are positive. Since none of the financial market imperfections has any impact, the analysis is equivalent to that in a world of perfect financial markets.

In examining case IV, it is possible to show that, when perfect financial markets are imposed on the model, fully anticipated inflation is neutral and the household will acquire assets in order to equate the rate of return on the house with the rate of return on financial savings. In case IV, the first-order condition (6) can be simplified to $P_H = P + V_H/\gamma$. Returning for the moment to the original notation, this is equivalent to

$$(12)\quad P_H = P_{HT}e^{-rT} + \int_0^T \gamma \frac{U_{At}}{U_{Bt}} e^{-rt}\,dt$$

To show neutrality, it is sufficient to show that the first-order condition and the budget constraint do not change when all prices are multiplied by e^{it} and all interest rates are replaced with $(r+i)$, where i is the rate of inflation. This is the case in equation (12),

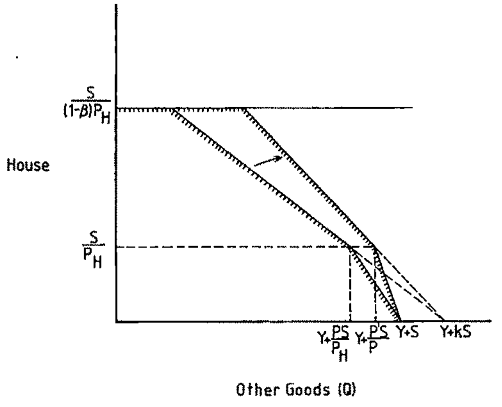


FIGURE 3

and can also be seen for the budget constraint by expressing it in the original notation.

To show that rates of return are equated across assets, first note that if the consumer were to purchase housing services in a rental market, utility maximization would require that $U_{Bt}/P_t = U_{At}/R_t$, where R_t is the rental price for one unit of house. Thus we can define R_t^* as the shadow price of one unit of house. Substituting this into (12) yields

$$(13) \quad P_H = P_{HT}e^{-rt} + \int_0^T \gamma R_t^* e^{-rt}\, dt$$

The consumer chooses the house size so that the price of the house at time 0 is equal to the price at T plus the sum of the shadow prices of housing services consumed from one unit of house from 0 to T, when all prices are discounted by r, the rate of return on the alternative asset.

This relationship continues to hold with an increase in P. The binding budget constraint for case IV is $S+Y=(P_H-P)H+Q$. An increase in P is a decrease in the price of housing services. A case IV consumer increases the size of house purchased, thus decreasing the shadow price of housing services, until equality in equation (13) is again attained. The choice of Q may increase or decrease.

Now consider a market model with a fixed supply of houses and a flexible price of houses. If all consumers are in case IV, equilibrium is attained when the net impact of changes in P and P_H on the demand for H is zero for each consumer. That is, equilibrium requires that $dP_H = dP$. Here the shadow price of housing services remains constant for each consumer even though the price of houses has increased. Equality is attained in (13) through the adjustment of P_H. Thus, with perfect financial markets an increase in the future price of houses results in (a) an increase in the initial price of houses equal to the present value of the future increase; (b) an increase in the percentage change in the price of house from 0 to T if P_H is greater than P; (c) no change in the paths of consumption or utility; and (d) a decrease in financial savings from 0 to T.

In case II, the mortgage market is used, but since the down-payment constraint is not binding the analysis of case II is similar to case IV. The binding budget constraint is $Y+kS=(kP_H-P)H+Q$. An increase in P results in an increase in the demand for H but an ambiguous change in Q. Since $S=(1-b)P_HH$, the increase in H implies that the fraction of the house mortgaged increases. With a fixed supply of houses and only case II consumers, equilibrium requires that $dP_H = dP/k$. The increase in P_H is less than the present value of the future increase because on the margin the consumer must finance the purchase with a mortgage.

In case III, the consumer holds no mortgage, but the stock of savings immediately after the house purchase is equal to zero. The value of the house purchase is set by the initial stock of savings, or $S=P_HH$. Thus an increase in P has no impact on the demand for H, but increases the consumption of Q by the amount HdP. In the market model with only case III consumers, an increase in the future price of houses has no impact on the initial price in equilibrium.

In case I, the downpayment constraint is binding, and the value of the house purchased is determined by the initial stock of savings and the downpayment constraint: $P_HH = S/(1-\beta)$. As in case III, an increase in P results in no change in the size of house purchased. Since $Q=Y+PH-k\beta P_HH$, an increase in P allows the consumption of Q to increase. To obtain equilibrium in the market model with only case I consumers, an increase in the future price of houses must result in no change in the initial price.

In summary, the partial equilibrium impacts of an increase in P are an increase in house size for cases II and IV, and no change in house size for I and III. In the market model for each case alone, there is no change in P_H in cases I and III, $dP_H = dP$ in case IV, and $dP_H = dP/k$ in case II.

B. *Equilibrium in a Market with Varying Initial Wealth Levels*

Here it is assumed that all consumers are identical in preferences and labor income streams, but differ in initial wealth levels. We assume $P < P_H$, and that there are some consumers in each of the four cases. No consumers enter or leave the market. Assuming that the housing stock is initially infinitely divisible, Part B analyzes the impact of a change in the future price of houses on the distribution of houses by initial wealth level and on the demand for mortgage funds.

If we impose an exogenous increase in the price of houses at the date of retirement, from Part A it can be seen that $0 < dP_H < dP$ in order for the housing market to be in equilibrium at time 0. By examining the budget constrains it can be seen that those in case IV, the highest wealth group, increase the size of the initial house purchase, and those in case I, the lowest wealth group, decrease their house sizes. If $k\ dP_H < dP$, then those in the second-lowest wealth group, case II, increase house size. House size decreases if the reverse inequality holds. In the second-highest wealth group, case III, house size unambiguously decreases.

To examine the impact of an increase in P on the mortgage market, consider a new market equilibrium where $0 < k\,dP_H < dP$. Then all case I and III consumers buy smaller houses, and all case II and IV consumers buy larger houses with an increase in P. Those on the margin between cases move from the higher to the lower-numbered case.[3]

The total demand for mortgage funds at time 0 is the sum over all consumers in cases I and II of bP_HH. In case I, $S = (1-\beta)P_HH$, so that P_HH does not change with the price increases. Thus the demand for mortgage funds by case I consumers remains constant. In case II, the fraction of the house mortgaged increases as does the value of the house purchased, so that the demand for mortgage funds increases for all case II consumers. Thus far it has been assumed that the supply of mortgage funds is perfectly elastic. The increased demand for mortgage funds by case II consumers is met, and some case III consumers may move to Case II, resulting in an increase in the number and average initial wealth of mortgage holders.

If the supply of mortgage funds is fixed, we can consider reallocating mortgage funds by either increasing the down-payment requirement or increasing the mortgage interest rate. The impact of decreasing β is to further decrease the size of houses purchased by the lowest wealth group. This dampens the increase in P_H, so that in comparison with perfectly elastic mortgage funds, rationing with a down-payment requirement results in smaller houses for the lowest wealth group and larger houses for all others. If instead the mortgage interest rate is increased, case II consumers decrease the demand for mortgage funds and the size of house purchased. This again dampens the increase in P_H so that there is a comparative increase in house size for cases I, III, and IV.[4]

C. *A Binding Liquidity Constraint at T*

The results in Section II are critically dependent on the assumption that the consumer can borrow perfectly against the value

[3] Those in case IV increase house size, decreasing the stock of savings immediately after the purchase. If this falls to zero, the household moves to case III. It can be shown that in case III the marginal rate of substitution of expenditure on H for expenditure on Q increases. If it increases to k, the household begins to mortgage part of the house and enters case II. In case II, b increases since $S = (1-b)P_HH$. If b increases to β the household enters case I.

[4] Holding P_H constant, a decrease in β may result in households moving from case II to I, while an increase in r_m may result in households moving from I to II and from II to III. The dampening of the increase in P_H, however, could result in case changes in either direction.

of the house at retirement. Now it is assumed that there is a liquidity constraint on financial savings at T. Let us consider the case where the stock of financial savings falls to zero at T, and retirement savings consists entirely of the sale value of the house.

In order to ease the analysis, $V(H,Q)$ is rewritten as the sum of two concave utility functions representing preretirement and retirement utility:

$$(14)\qquad V(H,Q)=M(H,G)+Z(N),$$

where

$$M(H,G)=\max\int_0^T U(\gamma H, B_t)\,dt$$

subject to $\quad G=\int_0^T P_tB_te^{-rt}$

where $N=We^{-rT}$ and $Z(N)=F(W)$. With S_T equal to zero, the consumer's maximization problem is

$$(15)\qquad \max M(H,G)+Z(N)$$

subject to
(a) $Y+S=(1-b+bk)P_HH+G$
(b) $N=PH$
(c) $0\leqslant b\leqslant\beta$
(d) $S\geqslant(1-b)P_HH$

With $S_T=0$, any further increase in house size increases the consumption of housing services and retirement assets, but at the sacrifice of the consumption of other goods before retirement. Thus if the liquidity constraint is binding at retirement, it is possible for all four cases to exist even with $P>P_H$. An analysis for case IV begins with the two budget constraints, $S+Y=P_HH+G$ and $N=P_H$. Assuming S, Y, and P_H constant, if G is chosen the level of preretirement utility is determined. Thus a preretirement utility function can be written as $M^*(G)$, where $M^*(G)=M((Y+S-G)/P_H,G)$. Substituting H out of the budget constraints, the lifetime maximization problem can be expressed as a simple maximization problem with a linear budget constraint:

$$(16)\qquad \max M^*(G)+Z(N)$$

subject to $\quad Y+S=G+P_HN/P$

An increase in P is a decrease in the price of retirement assets, so that N unambiguously increases. But, the change in G, and thus H, depends on the relative sizes of the income and substitution effects. Since an increase in P_H results in an unambiguous decrease in H, for a market equilibrium with identical case IV consumers, the present value of an increase in the retirement price is not necessarily equal to the increase in the initial price, and in fact could result in a decrease in P_H.

The results for case II are analogous, with the budget constraints written as $kS+Y=kP_HH+G$ and $N=PH$. Here S_T is always chosen equal to 0 if $P>kP_H$. An increase in P would result in an increase or decrease in H, so that market equilibrium with identical case II consumers could require an increase or decrease in P_H.

In cases I and III, the binding constraints completely determine the allocation of lifetime income to H, G, and N, so that an increase in P has no impact on the size of house purchased or the consumption of other goods before retirement. In all four cases the increase in P increases the value of retirement assets held.

D. *The Price Index for Retirement Assets*

In the preceding analysis it was assumed that the price index for retirement assets is invariant to changes in the price of houses at retirement, and prices normalized so that the price of retirement assets is equal to one. If, instead, the increase in the price of houses at retirement implies not only an increase in the dollar value of the house sold but an increase in the price of housing services purchased after retirement, the impact of an increase in P depends upon preferences for housing services after retirement.

Suppose that utility at retirement is a function of two goods, A and B. Let A represent postretirement housing services and have a price equal to P. Consumption of other goods after retirement are represented by B, with a price of one. At retirement the consumer maximizes a concave utility function, $X(A,B)$, with the constraint that total retirement wealth equal $PA+B$.

To examine the impact of an increase in P, we use the form of the lifetime utility function developed in Part C, but slightly alter the budget constraint. For case IV, the maximization problem is

$$(17) \qquad \max M(H,G)+Z(N)$$

$$\text{subject to} \quad S+Y=(P_H-P)H+G+P_N N$$

Define P_N as the cost of attaining $Z(N_0)$ with a new level of P relative to with P_0, where the subscript "0" denotes the initial value of a variable.[5]

Now consider the impact of an increase in P on the size of house purchase for case IV. An increase in P can be analyzed as a decrease in the price of houses and an increase in the price of retirement assets. There is a substitution effect towards H for both price changes, but the income effects have opposite signs. It can be shown that if the size of house purchased at time 0 exceeds the real values of expenditure on housing services after retirement, house size increases with an increase in P.[6] This result also holds for a case II consumer. For cases I and III the increase in P again results in no change in house size.

Assuming that $H_0>A_0$ for all consumers, the qualitative results for the market model presented in Part B are still valid. In a market with only case IV consumers, the market clears at time 0 when $0<dP_H<dP$. Denote this change in P_H as dP'_H. If the market with consumers in all four cases clears with $k\,dP'_H<dP$, then all case II and IV consumers purchase larger houses and all case I and III consumers purchase smaller houses.

III. Conclusion

A model of house-purchase decisions with imperfect capital markest has been developed, taking initial financial wealth, labor income, prices, and the purchase and sale dates of the house as exogenous. Results of the micro model and a simple market model have been discussed in analyzing the impact of an increase in the expected future price of houses. In the market model, redistribution of the housing stock due to changes in the price of houses and mortgage market conditions was discussed in relation to an increase in the future price of houses. A fixed supply of houses was assumed.

It is clear that changes in future house prices add to the difficulties in analyzing house and mortgage markets in the aggregate. Consumers respond very differently to changes in mortgage market conditions, even with identical preferences, depending on both the financial constraints that are binding and on the future price of houses. Also if changes in the future price of houses result in a redistribution of housing stock and mortgage funds, qualitative changes in market responses to current exogenous shocks in the housing and mortgage markets are implied.

REFERENCES

R. Artle and P. Varaiya, "Life Cycle Consumption and Homeownership," *J. Econ. Theory*, June 1978, *18*, 38–58.

W. Dolde, "Capital Markets and the Short-Run Behavior of Life Cycle Savers," *J. Finance*, May 1978, *33*, 413–28.

______ **and J. Tobin,** "Monetary and Fiscal Effects on Consumption," in *Consumer Spending and Monetary Policy: The Linkages*, Federal Reserve Bank of Boston, *Monetary Conference Series*, No. 5, Boston 1971.

Irving Fisher, *The Theory of Interest*, New York 1954.

D. Lessard and F. Modigliani, "Inflation and the Housing Market: Problems and Potential Solutions," in *New Mortgage Designs for Stable Housing in an Inflationary Environment*, Federal Reserve Bank of Boston, *Monetary Conference Series*, No. 14, Boston 1975.

F. Modigliani and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in

[5] Initially P_N is set equal to 1, so $P_N=(PA^*+B^*)/N_0$, where A^* and B^* are solutions to $\min(PA+B)$ subject to $X(A,B)=Z(N_0)$.

[6] The sum of the weights for the income effect is $H_0-N_0\,dP_N/dP$. Differentiating P_N with respect to P and using the first order conditions from the minimization problem in fn. 4 yields $dP_N/dP=A_0/N_0$. Thus the sum of the weights is (H_0-A_0).

Kenneth Kurihara, ed., *Post-Keynesian Economics*, New Brunswick 1954.

C. **Pissarides**, "Liquidity Considerations in the Theory of Consumption," *Quart. J. Econ.*, May 1978, *42*, 279–96.

W. **Poole**, "Housing Finance Under Inflationary Conditions," in *Ways to Moderate Fluctuations in Housing Construction*, Federal Reserve Staff Study, Washington 1972.

S. **Ranney**, "Capital Market Imperfections in a Model of House Purchase Decisions," unpublished doctoral dissertation, Univ. Wisconsin-Madison 1978.

T. **Russell**, "The Effect of Improvement in the Consumer Loan Market," *J. Econ. Theory*, Nov. 1974, *9*, 327–39.

L. **Thurow**, "The Optimal Lifetime Distribution of Consumption Expenditures," *Amer. Econ. Rev.*, June 1969, *59*, 324–30.

T. **Watkins**, "The Time Allocation of Consumption under Debt Limitations," *Southern Econ. J.*, July 1975, *42*, 61–69.

C. **Wiseman**, "Windfalls and Consumption Under a Borrowing Constraint," *Rev. Econ. Statist.*, May 1975, *57*, 180–84.

Vertical Integration: Does Product Price Rise or Fall?

By FRED M. WESTFIELD*

This article analyzes the effect on product price of vertical merger between an upstream monopoly seller of a raw material (input) and downstream competitive firms using this input, together with others, to produce a product. Popular wisdom is that such forward extension of monopoly power causes consumer prices to increase. This belief seems to underly new legislation in some states that prohibits oil companies, said to be monopolies, from owning filling stations, assumed to be competitors. It also provides support for proposals to prohibit vertical integration or to force vertical dismemberment.

Vertical integration has been often analyzed.[1] One strand of the literature, which will not be considered here, studies vertical integration of successive stages of imperfect markets. Removal, through merger, of a price-marginal cost wedge for an intermediate good lowers downstream costs and, therefore, downstream product price.[2] The problem considered in this paper comes from a different strand of the literature. I analyze vertical integration that leads to monopoly in a previously competitive market for the downstream product. Thus one effect of vertical merger is an increase in the price of the final product as a result of monopolization of that market and a new wedge between price and marginal cost. At the same time, the merger eliminates the wedge between the monopoly price of the input and its marginal cost because the merged firm charges itself the marginal cost. The effect of this is to reduce unit cost and, therefore, price of final product. And so one needs to determine which of the two effects on product price, acting in opposite directions, is dominant.

There is general agreement that the opposing effects exactly offset each other where inputs are utilized in fixed proportions by the downstream industry. Product price in this case is unaffected by a change in the vertical market structure. Ward Bowman, however, provided general arithmetical examples designed to show that product price could either increase or decrease. George Hay, Richard Schmalensee, and Frederick Warren-Boulton each argued that if a Cobb-Douglas production function represents the technological possibilities of producing the final commodity and demand for it has a constant elasticity, then vertical integration would have to result in an increase in the equilibrium price of output. Moreover, Hay and Warren-Boulton believed, incorrectly as it turns out, that the final product price would increase if technology is represented by any *CES* production function, for the entire spectrum of elasticities of substitution between zero and infinity. Warren-Boulton based his conclusions mainly on extensive computer simulations, while Hay did his own numerical experiments and examined Warren-Boulton's results. Schmalensee chose more general analytical techniques, but was unable to deduce the effect on product price. He conjectured that the outcome depends on more specific "...information about cost and demand conditions" (p. 446).

This paper develops a general analytical scheme based on production-cost duality. The method does not postulate specific mathematical forms of production or demand functions. The basic idea is to take the price-quantity equilibrium configuration of inte-

*Professor of economics, Vanderbilt University. I am indebted to my colleagues Ben Bolch and Elton Hinshaw for asking me the question that serves as title of this paper.

[1] For a valuable review of the vertical integration literature, see David Kaserman. Included is a bibliography of more than sixty items.

[2] An exception occurs if the downstream product is sold by monopoly and the upstream intermediate good is an inferior (i.e., regressive) input whose reduced "price" through merger raises downstream marginal cost.

grated monopoly as starting point. This equilibrium is determined by the demand function for the product and the costs for the product generated by a given production function and input prices. Next, consider the alternative nonintegrated vertical industry structure. Suppose that upstream monopoly tries a price for the input which raises downstream marginal costs to a level just equal to the profit-maximizing price of output charged by integrated monopoly. This price of the input, not necessarily a profit maximizing price for upstream monopoly, will be called the *benchmark* price of the input. A competitive downstream industry, required to purchase the input at the benchmark price, would then be in equilibrium at the identical price and quantity for output as is the integrated monopoly. The critical question is whether, at the benchmark price, upstream monopoly's marginal revenue, corresponding to *mutatis mutandis* derived demand, is greater or less than marginal cost. Ruling out multiple *MR-MC* intersections, I thus determine whether the upstream monopoly's profit-maximizing price is lower or higher than the benchmark price. For a profit-maximizing price lower (higher) than the benchmark price, downstream marginal and average cost are lower (higher) than integrated monopoly equilibrium price of final output. Therefore, a downstream competitive price, based on the upstream monopoly equilibrium price, is lower (higher) than integrated monopoly equilibrium price that is based on the marginal cost-price of the input.

The results are not quickly summarized. Evidently, the price elasticity of derived demand faced by upstream monopoly, evaluated at the benchmark price, is an important consideration. If inelastic (less than one), marginal revenue for the quantity of input supplied is negative and surely less than marginal cost. The competitive equilibrium price for the product, combined with an upstream monopoly, ends up higher than the profit-maximizing price for the product of an integrated monopoly. The analysis is simplified when one assumes a linear homogeneous production function and also horizontal supply curves for the inputs. Then elasticity of derived demand for the input, evaluated at the benchmark price, is the weighted mean of price elasticity of demand for integrated monopoly in equilibrium and of the elasticity of substitution evaluated at the benchmark price, the weights being "relative shares." As the price elasticity for the product must be greater than one, the elasticity of derived demand for input faced by upstream monopoly, given the weights, is higher than price elasticity of demand and elastic or lower and possibly inelastic, according to whether the elasticity of substitution is higher or lower than the elasticity of demand for the product.

I show that if upstream monopoly charges the benchmark price and elasticity of substitution is so much greater than unity so as to exceed the elasticity of demand for product at the integrated monopoly equilibrium price, upstream monopoly's marginal revenue will exceed marginal cost. Vertical integration would increase product price. But this outcome may occur also if a nonzero elasticity of substitution is less than the elasticity of demand for product. For a Cobb-Douglas (constant-unitary-elasticity-of-substitution) production function, for example, vertical integration increases price regardless of how low the price elasticity of (elastic) demand, even for non-isoelastic demand curves. Vertical integration, moreover, raises the product price when *complete* substitution against the monopolized input would occur at a hypothetical input price for which downstream unit costs are below the integrated monopolist's equilibrium price. Though a benchmark price is here undefined, *every* possible downstream competitive price is then lower than the monopoly price, regardless of the price of the monopolized input. However, if a nonzero elasticity of substitution evaluated at the benchmark price is less than elasticity of demand for integrated monopoly equilibrium, then upstream monopoly's marginal revenue can also lie below marginal cost, and vertical integration will lead to a reduced product price. The cost saving made possible through integration more than compensates for the effect of monopoly power in the product market. Indeed, contrary to Hay's and Warren-Boulton's belief, this occurs for *CES* production functions, when the elasticity of substitution is less than unity and price elasticity of demand, greater than one, is low.

Although vertical merger may lead to product price decrease when a nonzero elasticity of substitution at the benchmark price is low, a sufficiently high price elasticity of product demand for integrated monopoly, producing given equilibrium output, will always outweigh this effect of low substitution elasticity.

For an elasticity of substitution at the benchmark price equal to zero, vertical merger will not cause a product price rise. Product price will definitely fall if the merged firm does substitute to some degree in favor of the otherwise monopolized input, even though, for a price of input as high as the benchmark price, the substitution elasticity is zero at an isoquant corner (ridge line) which prohibits further substitution against the input. However, in the extreme case that includes *CES*, where no substitution among inputs occurs as a result of the changed effective relative input costs following merger—zero elasticity of substitution not merely at the benchmark price but over the entire relevant range of input prices—vertical merger leaves product price unaffected.

I. The Model

The quantity of final product sold is designated by q. Input quantities required are x and y, and the production function for final product is assumed linear-homogeneous

$$q=F(x,y) \tag{1}$$

so that the unit-isoquant is given by

$$1=F(x/q,y/q)=F(a,b) \tag{2}$$

Prices for the inputs are assumed to be parametric from the point of view of the purchaser. Consequently, given the price w for x and s for y, cost minimization by producer(s) of final product, together with the homogeneity assumption, determines the input coefficients:

$$a=a(s/w),\, b=b(s/w) \tag{3}$$

as functions of relative input prices.

$$a'\geqslant 0,\quad b'\leqslant 0 \tag{4}$$

because of convex-to-the-origin isoquants. The equalities of (4) remind us that increased relative prices of inputs may at isoquant "kinks" or "corners" result in unchanged input coefficients.

Consider the input y as resource or intermediate good. Its unit cost of production is r, independent of the amount produced or of the vertical structure. Without vertical integration the producers of the final product q cannot also enter the industry producing y. They must purchase at the price s from a monopolist who obtains it at the constant unit cost r. With vertical integration, the unit cost of y to the combined firm is also r, so that, in effect, $s=r$. The other input, x, is sold at the constant price w, whether production is vertically integrated or not.

Under these assumptions average and marginal cost of production, λ, do not depend on quantity of final product q. For given prices s of y and w of x

$$\lambda=wa(s/w)+sb(s/w)=\lambda(w,s) \tag{5}$$

where a and b satisfy a generalized cost-minimizing isoquant-isocost tangency[3]

$$wa'+sb'=0 \tag{6}$$

Demand for final product is

$$q=D(p),\quad D'<0 \tag{7}$$

where p is the price of final product.[4] Inverted, this function gives average revenue $p=H(q)$ and total revenue

$$R=H(q)q=R(q) \tag{8}$$

[3] If the isoquant has a kink or corner, isoquant and isocost touch without crossing: $a'=b'=0$.

[4] If final product is potentially supplied by a segment of the industry not at all dependent on the input y (by

Marginal revenue, R', is assumed positive for sufficiently small q, that is, demand is price elastic at such a point, and the slope $R''<0$.

A. *Integrated Monopoly*

If the industry is vertically integrated the price s of the input y is its cost r. Designating by subscript "0" the equilibrium magnitudes resulting from this organization of production, marginal and average cost of final product is

$$(9)\qquad \lambda_0=a_0w+b_0r=\lambda(w,r)$$

The input-output ratios a_0 and b_0, given by (3) with $s=r$, satisfy the cost-minimization condition (6). It is assumed that integrated monopoly uses some of both inputs: $a_0, b_0>0$. The profit maximizing output q_0 is determined by the marginal revenue-marginal cost equality

$$(10)\qquad R'_0=R'(q_0)=\lambda_0$$

and monopoly equilibrium price p_0, by the average revenue function or from

$$(11)\qquad p_0=R'_0E_0/(E_0-1)$$

where $E_0=-d\ln q/d\ln p$, the value of the price elasticity of demand evaluated at q_0, a positive number greater than unity.

All this takes for granted that the monopoly has economic value—that λ_0 is sufficiently low relative to the vertical intercept of the demand curve so that $q_0>0$.

B. *Nonintegrated Upstream Monopoly*

The alternative vertical organization of the industry involves pure competition in the final product market by firms selling q and purchasing their input requirements x and y as price takers. Price p for the product is determined by market clearing at zero-profit

$$(12)\quad p(q)=\lambda(w,s)\equiv wa(s/w)+sb(s/w)$$

Price w is unchanged, but the price s of y is now set by the upstream monopolist maximizing his profit

$$(13)\qquad \phi=(s-r)bq$$

As s is raised by the monopoly, downstream competitive firms substitute against the input y, if technology allows, lowering b, raising a. As long as substitution against y is not complete so that b is positive, increases in b shift marginal cost λ upward. That is,

$$(14)\quad d\lambda/ds=a'+(s/w)b'+b=b>0$$

With downstream marginal cost higher because of increased s, the competitive market-clearing price given by (12) must increase and the quantity q of output demanded will decrease according to the demand curve for final product. That is, by differentiating (12) which, for given w, is an implicit equation in q and s, we obtain

$$(15)\qquad dq/ds=(d\lambda/ds)/p'=b/p'$$

If substitution against y is total, $d\lambda/ds=b=0$ and $dq/ds=0$.

The rate of change in upstream monopoly profit in response to change in price s is examined by differentiating (13) with respect to s:

$$(16)\quad d\phi/ds=(s-r)(b\,dq/ds+q\,db/ds)+bq$$

As input proportions adjust for least cost and final product price increases as it must, one obtains, substituting for dq/ds from (15),

$$(17)\qquad d\phi/ds=bq\left\{1-\frac{s-r}{s}\left[\frac{bs}{p}\cdot\frac{-p}{p'q}+\frac{-db}{ds}\cdot\frac{s}{b}\right]\right\}$$

or

$$(18)\quad d\phi/ds=bq\{1-[(s-r)/s][\beta E+e]\}$$

where bq is the quantity of y sold; $(s-r)/s$

an entirely different technology), then $D(p)$, the demand function for final product, is to be thought of as an excess-demand function and q, that quantity of final product *not* supplied, at each price, from this alternate source of production. Thus the demand curve for final product might be bounded by a limit price.

is the profit per unit of y expressed as percentage of the selling price; $bs/p=\beta$ is the "relative share" of the input y, also equal to $1-\alpha$, where $\alpha=wa/p$ is the relative share of x; E is the price elasticity of final demand; and $e=-(db/ds)(s/b)$ is the price elasticity of demand for input y, final output q held constant—of the output-compensated demand for y. It can be shown that $e=\alpha\sigma$, where σ is the elasticity of substitution.

From (18), we see that if the upstream monopolist of y sells anything at all so that $bq=y>0$,

(19)
$$d\phi/ds \gtreqless 0, \text{ as } N=[(s-r)/s][\beta E+e] \lesseqgtr 1$$

For upstream monopoly profit to take on its maximum value, call it ϕ_1, $d\phi/ds=0$. The number N, the center of attention below, will equal unity. This is the first-order condition for the interior maximum. It determines the price s_1 for the quantity y_1, the input coefficients a_1 and b_1, and the downstream price and quantity p_1 and q_1. The subscript "1" indicates equilibrium values when production is not integrated—upstream monopoly maximizing profit, downstream competitive firms combining inputs efficiently, product price equaling cost and market for product cleared. The condition is equivalent to the requirement that (derived) marginal revenue, $MR=s(1-1/\eta)$, marginal to the *mutatis mutandis* derived demand curve (AR), product sold competitively, is equal to marginal cost, $MC=r$, of y. This follows at once from the fact that $\beta E+e=\beta E+\alpha\sigma$ is the Hicksian price elasticity η of derived demand for y (see Roy G. D. Allen, p. 373).

II. The Comparison

If an upstream monopolist has any market power at all ($\eta_1<\infty$), equilibrium price s_1 exceeds cost r. Unit cost of producing final product must obey the inequalities

(20)
$$\lambda_1=wa_1+s_1b_1>wa_1+rb_1\geqslant wa_0+rb_0=\lambda_0$$

and input coefficients, the inequalities $a_1\geqslant a_0$, $b_1\leqslant b_0$ (convexity). The first inequality in (20) holds because $s_1>r$ and the second because a_0 and b_0 are cost-minimizing values for the input coefficients at input prices w and r (equation (9)). If the isoquant is smooth in addition to being convex so that $a(s/w)$ and $b(s/w)$ have continuous first derivatives, the second is a strict inequality because of *some* substitution against input $y(a_1>a_0, b_1<b_0)$ in response to its increased price. This might be termed the McKenzie-Burstein-Vernon and Graham substitution effect, the consequence of market imperfection in the vertical structure.[5] If the isoquant has a kink at a_0, b_0 the substitution effect, however, may not occur; the second weak inequality is then satisfied as equality. Thus for an activity-analysis isoquant with kink at a_0, b_0 between facets or convex segments of the isoquant, the equality *may* hold because s_1 is not sufficiently large for substitution to be cost effective. For the Leontief-fixed coefficient isoquant, it *must* hold because substitution is impossible.

Maximum profits Π_0, achievable by monopoly under vertical integration, cannot be less than maximum profit ϕ_1, achievable without it. This is deduced from the fact that Π_0 maximizes $(p\text{-}wa\text{-}rb)q$ at q_0, a_0, b_0, with w and r as parameters, while ϕ_1 is a value of the same function evaluated at q_1, a_1, b_1. In fact,

$$\begin{aligned}\Pi_0&=[p(q_0)-wa_0-rb_0]q_0\\&\geqslant[p(q_1)-wa_0-rb_0]q_1\\&\geqslant[p(q_1)-wa_1-rb_1]q_1=\phi_1\end{aligned}$$

equalities holding if substitution among inputs does not occur because it is technologically impossible or else uneconomical.

The effect of altered vertical structure on final product price and quantity is more complex. Consider the output q_0 sold at price p_0 by the vertically integrated profit-maximizing monopoly as a *benchmark*. If the

[5] Named after writers who illuminated it in 1951, 1960, and 1971, respectively.

same quantity and price are to prevail in the final product market in the absence of vertical integration, then the upstream monopolist would have to charge a price s^* so that

$$(21)\quad p_0 = \lambda^* = wa^*(s^*/w) + s^*b^*(s^*/w)$$

This price s^*, if it can exist for a positive quantity y^*, would induce downstream competitive purchasers to choose the least-cost input mix a^*, b^*, and it would impose on them a marginal and average cost λ^* equal to p_0. The price s^* is called the benchmark price of y and λ^*, the benchmark marginal cost. The value of any variable of the model evaluated at s^* and downstream market for product in competitive equilibrium is called the benchmark value and is designated by an asterisk. Thus, from the definition, $p^* = p_0$ and $q^* = q_0$.

Generally the benchmark price s^* is not equal to the upstream monopoly's profit-maximizing price s_1, though it could be. One needs to see whether $s_1 \gtreqless s^*$ to determine whether $p_1 \gtreqless p_0$ (and $q_1 \lesseqgtr q_0$). Downstream marginal cost λ_1, to which p_1 is equated by competitive market forces varies directly with s_1 as seen in (14) and (20). Therefore, if $s_1 \geqslant s^*$ then $\lambda_1 = p_1 \geqslant \lambda^* = p_0$, and $q_1 \leqslant q_0$; if $s_1 < s^*$, then $p_1 < p_0$ and $q_1 > q_0$.

But how to discover the relationship between s^* and s_1? With price s^* of y and the market for final product cleared, evaluate the upstream monopolist's $MR^* = s^*(1 - 1/\eta^*)$. If this benchmark marginal revenue exceeds MC^* or r, then $y_1 (= b_1 q_1) \geqslant y^*$; and the profit maximization by the upstream monopolist entails $s_1 < s^*$. To increase profit, upstream monopoly moves down along the *mutatis mutandis* derived demand and marginal revenue curves for y, with downstream marginal cost and final product price lowered from their benchmark levels by market forces as s is lowered by the monopoly. The upstream profit-maximizing position thus requires a price p_1, lower than the benchmark price p_0, required for profit-maximizing integrated monopoly. Conversely, if evaluation at the benchmark, indicates $MR^* < r$, then price s_1 exceeds s^*, and p_1 exceeds p_0. This argument takes for granted that the MR curve, marginal to the negatively sloped derived demand curve for y (pure competition in the market for final product q), has negative slope—elasticity η of derived demand does not increase with reductions in price s so fast as to cause marginal revenue to be an increasing function of the quantity y in the relevant neighborhood of y^*—ruling out the possibility of multiple $MC = MR$ intersections.

The price s^* satisfying (21) may not exist. Suppose that isoquants intersect the x-axis and that complete substitution against y is economical for prices s higher than $\bar{s}$. That is for $s \geqslant \bar{s}$, $\bar{b} = b(s/w) = 0$. Then the downstream marginal cost attains an upper bound $\bar{\lambda} = \bar{s}b(\bar{s}/w) + wa(\bar{s}/b) = w\bar{a}$. As upstream monopolist attempts to increase price, demand for y goes to zero at $\bar{s}$ regardless of the amount of final output produced downstream; and if this occurs for $\bar{\lambda} < p_0$, the benchmark price s^* is not defined. But in this situation, discussed in Section III, Part A below, the competitive product price p_1, in the absence of integration, is necessarily lower than the monopoly price p_0, with integration.

If the benchmark price s^* does exist, $sgn(MR^* - MC^*) = sgn(-d\phi^*/ds)$ is evaluated by calculating at the benchmark the number N defined by (19). Thus we have a relationship, all important for our analysis, between

$$(22)\qquad N^* = [(s^* - r)/s^*][\eta^*]$$

where $\eta^* = \beta^* E_0 + \alpha^*\sigma^*$ and $\alpha^*\sigma^* = e^*$, and the two equilibrium positions. That is,

$$(23)\quad N^* \gtreqless 1,\ s^*(1 - 1/\eta^*) - r \gtreqless 0,$$
$$s^* \gtreqless s_1,\ p_0 \gtreqless p_1,\ q_0 \lesseqgtr q_1$$

I call N^* the *critical benchmark number*.

The critical benchmark number is clearly less than unity for elasticity η^* of derived demand less than unity. Monopoly price s_1 then surely exceeds benchmark price s^* and product price p_1 sans integration exceeds product price p_0 with integration. As η^* is the weighted mean of product demand elasticity ($E_0 > 1$) and substitution elasticity σ^*, such inelastic derived demand requires $\sigma^* < 1$. But, of course, $\sigma^* < 1$ is not sufficient for $\eta^* < 1$; and $\eta^* < 1$ is not necessary for $N^* < 1$.

Properties of the production and demand functions for final product determine the critical benchmark number N^*. However, the relevant production function properties are, because of duality, captured in the least-cost function $\lambda(w,s)$.[6] It is convenient to simplify notation by setting $w=1$, so as to measure the input x in terms of dollars' worth, and writing

$$(24) \qquad c=\lambda(1,s)=c(s)$$

I call $c(s)$ the marginal cost frontier (MCF).[7] From (14), $c'(s)=b(s)\geqslant 0$ and from (4), $c''(s)=b'(s)\leqslant 0$. Thus, the components of η^* are

$$\beta^* E_0=\frac{s^*b(s^*)}{p_0}\cdot\frac{p_0}{\lambda^*-\lambda_0}=\frac{s^*c'(s^*)}{c(s^*)-c(r)}$$

$$=\left\{\frac{d\,ln[c(s^*)-c(r)]}{d\,ln(s^*-r)}\right\}\left\{\frac{s^*}{s^*-r}\right\}$$

$$\alpha^*\sigma^*=e^*=\frac{-s^*c''(s^*)}{c'(s^*)}=\frac{-d\,ln\,b(s^*)}{d\,ln\,s^*}$$

and

$$(25) \quad N^*=\frac{(s^*-r)c'^*}{c^*-c_0}+\frac{-(s^*-r)c''^*}{c'^*}$$

or

$$(26) \quad N^*=1+\left[\frac{-d\,ln\,c'^*}{d\,ln(s^*-r)} - \frac{-d\,ln[(c^*-c_0)/(s^*-r)]}{d\,ln(s^*-r)}\right]$$

The difference of (negative) logarithmic derivatives in (26) will, in a moment, be interpreted as a difference between two elasticities. Meanwhile, as $1+d[ln(c^*-c_0)/(s^*-r)]/d\,ln(s^*-r)=d\,ln(c^*-c_0)/d\,ln(s^*-r)$, we have an interpretation, not entirely transparent, that the critical benchmark number is greater or smaller than one according to whether a 1 percent incremental increase in the markup s^*-r, at the benchmark, on y^* sold by the nonintegrated upstream firm, would reduce the input-output ratio $b(s^*)=c'(s^*)$ by a percentage greater in absolute value than it would increase the percentage in the markup $c^*-c_0=p_0-c_0$ on q_0 sold by the integrated monopoly.

It is also useful to note that

$$(27) \quad N^*=1-\frac{c^*-c_0}{c'^*}\left[\frac{d}{ds}\,\frac{(s^*-r)c'^*}{c^*-c_0}\right]$$

which can be verified by carrying out the differentiation.

Figure 1 clarifies. In Panel B is a sketch of MCF with c on the vertical and s on the horizontal.[8] Because of convex isoquants, the curve is concave. Its c-intercept is $a(0)$: equal to zero where technology allows complete replacement of x when y is free (the isoquants intercepting the y-axis), or else equal to the minimum quantity x, in dollars' worth, required per unit of final output even if y is free (isoquants never touching the y-axis). The slope c' of MCF at s is the input coefficient $b(s)$, and the c-intercept of the tangent at s is the input coefficient $a(s)$. In Panel C is plotted the curve "marginal" to the "total" or MCF curve. It is the input coefficient $b(s)$, proportional to the output-compensated demand curve for y, with price on the horizontal, quantity on the vertical; and if one normalizes $q_0=q^*=1$, it actually is the demand curve for y, final product held constant at the integrated monopoly equilibrium level—at benchmark output. The slope $b'(s)$ of this demand curve is, of course, nonpositive because of the concavity of MCF. The output-constant demand curve could have a horizontal segment, $c''=b'=0$, an isoquant's kink being reflected in linearity of MCF. Note that if technology calls for the complete substitution of x against y at a price $\bar{s}$, for which the isoquant intersects the x-axis, and, therefore, also for $s>\bar{s}$, then, for $s\geqslant\bar{s}$, MCF

[6] Schmalensee also used cost-production duality relationships in his investigation.

[7] The name is suggested by Paul Samuelson's coinage of "factor price frontier" defined by the implicit function $1-\lambda(w/\lambda, s/\lambda)=0$.

[8] The interested reader may wish to compare the properties of the standard total product curve given by $q=F(1,y)$ with corresponding properties of MCF, to fully appreciate the duality.

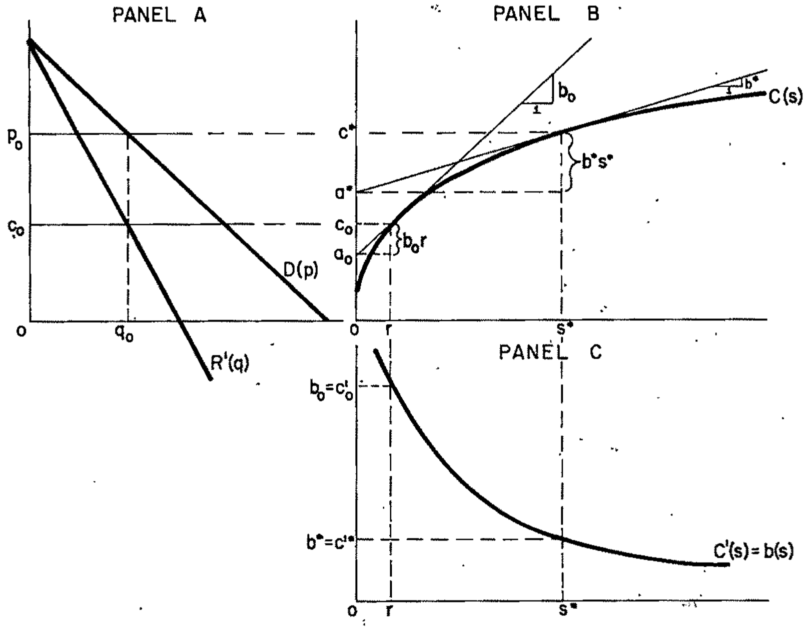


FIGURE 1. PANEL A: DEMAND [$D(p)$], MARGINAL REVENUE [$R'(q)$] AND INTEGRATED MONOPOLY'S UNIT COST [c_0] CURVES; PANEL B: THE MARGINAL COST FRONTIER [$c(s)$]; PANEL C: THE CURVE MARGINAL TO THE FRONTIER, SHOWING OUTPUT-COMPENSATED DEMAND.

is horizontal at level $a(\bar{s})$; the output-constant demand $c'(s)=b(s)$ for y in Panel C, thus zero, then coincides with the horizontal (price) axis.[9] If, on the other hand, isoquants never do touch the x-axis, MCF is unbounded, and the demand $b(s)$ ultimately approaches a constant as s is increased. The demand elasticity e approaches zero.

Panel A of Figure 1 shows demand for final output, the conventional marginal revenue curve for monopoly of final output, and the integrated monopoly's (horizontal) marginal and average cost curve at the level $c_0=a(r)+rb(r)$, as cost of y is assumed to be r.

Both the integrated monopoly equilibrium and the benchmark configurations are completely depicted on the diagram. Final product price p_0 and output q_0 for integrated monopoly are given in Panel A by the intersection of $R'(q)$ and c_0. The quantity (cost) a_0 of the input x is given on the c-axis by the intercept of the tangent to MCF at r; the quantity b_0 of input y, by its slope and also by the ordinate at r in Panel C. The benchmark price s^* of y is seen to be determined by the abscissa in Panel B corresponding to the benchmark marginal cost c^* at price p_0 on the ordinate. The benchmark values of the input coefficients a^* and b^* are, respectively, given by the intercept and slope of the tangent to MCF at s^*; and b^* is also the ordinate in Panel C evaluated at s^*.

Whether the critical benchmark number N^* is greater or less than one can now be easily determined graphically. Figure 2 replicates the marginal curve, the output-compensated demand, from Panel C of Figure 1. It also shows an "average" curve $(c-c_0)/(s-r)$ derived in textbook fashion from the "total" MCF-curve, with r, c_0 serving as origin. This average curve gives the ratio of increased minimum cost for downstream product $(c-c_0)$, to the increased cost for the intermediate product, $(s-r)$. The average

[9] The complete substitution of y for x at a nonzero price of y (intersection of isoquants with the y-axis) implies a linear MCF through the origin for prices between zero and such a price. As it is assumed that integrated monopoly uses some of both inputs, r and s^* are higher prices than this.

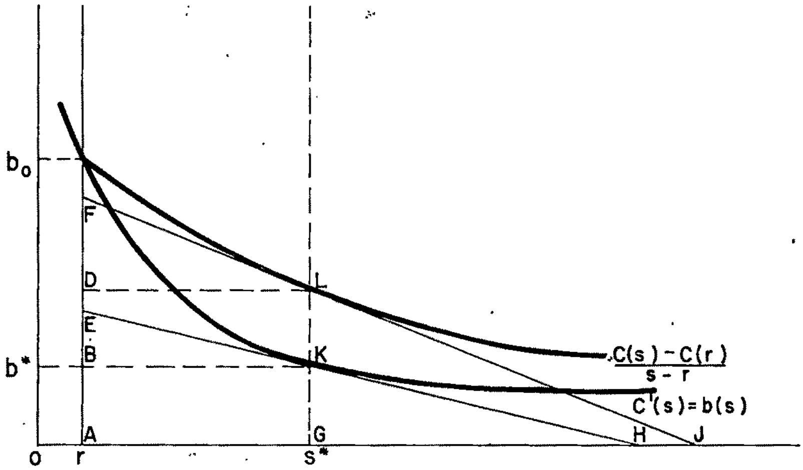


FIGURE 2. MARGINAL CURVE (FIGURE 1, PANEL C) SHOWN WITH CORRESPONDING AVERAGE CURVE

and marginal curves must, of course, coincide as $s \to r$ because $lim(c-c_0)/(s-r)=c'(r)$. Further, $(c-c_0)/(s-r) \geqslant c'(s)$ for $s>r$, and neither function may increase with s because of concavity of *MCF*—convexity of isoquants.[10] The elasticity at a benchmark price s^* of the average curve, with (r, c_0) as origin, is $-d\,ln(c^*-c_0)/(s^*-r)/d\,ln(s^*-r)$ or AG/GJ in Figure 2, while the elasticity of the marginal curve is $-d\,ln\,c'^*/d\,ln(s^*-r)$ or AG/GH. The formula (26) then tells us at once that if at s^* the elasticity of the marginal curve, AG/GH, exceeds the elasticity of the average curve, AG/GJ, $N^*>1$; if not, $N^* \leqslant 1$. In short:

$$N^* \gtreqless 1 \text{ according as } GJ \gtreqless GH$$

If the tangents at s^* to the average and marginal curves intersect for $s>r$ above the s-axis $N^*<1$; if not, $N^*>1$. If they intersect on the axes, the two elasticities are equal and $N^*=1$.

To sort out how final demand, technology and input costs interact to determine the outcome of vertical integration, observe how N^* changes in response to exogenous shifts in relevant parameters. For final demand the critical variable is the price elasticity $E_0 \equiv p_0/(p_0-c_0)$, evaluated at integrated equilibrium (benchmark) output. For example, the effect of a more price-elastic final demand, given c_0 and q_0, is studied by shifting $D(p)$ so as to rotate the marginal revenue curve $R'(q_0)$ counterclockwise about the coordinate q_0, c_0 in Panel A of Figure 1.[11] Thus, $p_0=c^*$, together with s^*, are reduced while c_0, r and the position of the curves in Figure 2 are unchanged. Indeed, N^* can be considered a function of s^*: the decreases of c^* and s^* showing the effects on N^* (given the production function and the costs r and w) of increased elasticity at given output q_0.

Different production functions, with c_0, w, r and demand for the final product held constant, on the other hand, change the shape of the *MCF* and, hence, the relevant elasticities for the average and marginal curves of Figure 2. Accordingly, N^* is a function also of parameters characterizing technology.

One may investigate effects of changes of input costs w and r. The analysis of this paper treats them as parameters—supply curves for inputs are horizontal. Alternatively, input prices might be systematically related to quantities purchased by the industry and, to market power exercised by purchasers. Warren-Boulton (pp. 796–99) considered the supply price w increasing

[10] Cost minimization and convexity imply $a(r)+rb(r) \equiv c_0 \leqslant a(s)+rb(s) \equiv c(s)-(s-r)b(s)$. Or $c_0 \leqslant c(s)-c'(s)(s-r)$.

[11] As q_0 and p_0 are determined at integrated monopoly equilibrium, $1<E_0<\infty$, regardless of how high or low p_0 is relative to c_0.

along a constant-elasticity supply curve for x. If, regardless of vertical structure, no monopsony power is exercised by purchasers, this complication requires that η^* in formula (22) for critical benchmark number be modified (see J. R. Hicks, p. 244) to include an effect for the supply elasticity of x. A competitively rising supply price of y would require that r be replaced by an increasing supply function of y. On the other hand, constant degree of monopsony power for y would require replacement of r by a marginal cost function that involves also the elasticity of this supply curve. In general, supply of each input is a function of both input prices w and r; and equilibrium is influenced by likely changes in degree of monopsony power that may result from reduction through vertical integration in the number of independent purchasers of the input x.

The analysis ignores possible shifts of the production function as the result of merger. Output produced from any given input combination x, y could increase or decrease. An increase may occur because of an improvement in the coordination of production; a decrease, because of inability of a conglomerate to replicate efficiencies of decentralization. Such technological externalities strengthen, respectively weaken, the cost and price reduction effects of vertical merger.

III. Some Cases

To conclude the analysis, I now apply my method to study how added restrictions on the forms and characteristics of production functions sharpen the conclusions about consequences of integration. I deduce the specific results summarized in the introduction and obtain several additional insights which could not be easily summarized at the beginning.

A. $c(s)$ *is Bounded by* $c(s)=\bar{a}$

This is a production function for which isoquants *intersect* the x-axis. A ridge line coincides with the x-axis, and it is technically possible to produce q without at all utilizing y. Thus, the unit isoquant intersects the x-axis at, say, $\bar{a}=a(s)$, and $a'(s)=b'(s)=0$, for $s \geqslant \bar{s} > r$.

From the geometry, if $p_0 > \bar{a}$, the benchmark is undefined, and if product demand at integrated monopoly has elasticity $E_0 = p_0/(p_0-c_0) \geqslant \bar{a}/(\bar{a}-c_0)$, any downstream competitive price, equal to $c(s)$, is less than p_0. Consequently, if isoquants intersect the x-axis, a sufficiently low price elasticity of demand for the product will always insure that integration raises price.

For *CES* production functions, $1<\sigma\leqslant\infty$, Kenneth Arrow et al. show that isoquants must intersect the axes. Therefore, for such production functions and sufficiently low E_0, product price must surely rise. (Also see Part D below.)

B. $b''(s)\leqslant 0$ *for* $b(s)>0, s>r$

This means that convex isoquants for final product are such that output-compensated demand $b(s)$ is linear or concave (from below). In that case we see that $p_0>p_1$. Geometrically, the tangents to the marginal $b(s)=c'(s)$ and the average $(c(s)-c_0)/(s-r)$, given r, do not intersect to the right of the price s^* determined by value of the price elasticity E_0 if the benchmark is defined. And, as we just saw, for low values of E_0 for which the benchmark is not defined, $p_0>p_1$ also.

Where a benchmark is defined, the result is proven analytically. Note first that with q_0, r, w and c_0 fixed, as $E_0=c^*/(c^*-c_0)\rightarrow\infty$, from (25),

$$(28)\quad \lim_{s^*\rightarrow r} N^*(s^*)$$

$$= \lim_{s^*\rightarrow r}\left[\frac{c'^*(s^*-r)}{c^*-c_0}+\frac{-c''^*(s^*-r)}{c'^*}\right]$$

$$=1+0=1$$

Now, set $p_0=c^*$ and allow it to increase with s^* so as to study the effect of lower price elasticity on the magnitude of N^*. Then

$$\frac{dN}{ds}=\frac{N}{s-r}\left[1-\frac{c'(s-r)}{c-c_0}\right]$$

$$+\frac{c''^2}{c'^2}(s-r)-\frac{c'''(s-r)}{c'}$$

where asterisks are left off to simplify notation. N^* is positive (being unity in the limit $s^*\to r$) for $s^*>r$, and $c'^*(s^*-r)/(c^*-c_0)\leq 1$ (see fn. 10). Hence, for $c'''^*\leq 0$, which is the condition for linear or concave-from-below output-compensated demand of y, $dN^*/ds>0$ and $N^*>1$. Regardless of how high or low the elasticity of demand for final product, vertical integration raises price of final product.

It can be shown that for the *CES* family of production functions, $\sigma>0$, the output-compensated demand curve is, in fact, strictly convex from below. Thus $c^{*'''}\leq 0$ requires variable elasticity of substitution, σ increasing with s.

C. *Constant e*

For constant price elasticity of output-compensated demand, the linear homogeneous production function is of the Cobb-Douglas form: that is, $\sigma\equiv 1$, and $e=\alpha=1-\beta$ is the constant relative share for the input x.[12] In this case, vertical integration raises price.

Diagrammatically, the tangents of Figure 2 would not intersect to the right of s^* for a Cobb-Douglas *MCF*. More directly, substituting in (25),

$$N^*=\left(\frac{s-r}{s}\right)\left(\frac{\beta s^{\beta}}{s^{\beta}-r^{\beta}}+(1-\beta)\right)$$

where asterisks are again for simplicity suppressed. To show $N^*>1$ for $s^*>r$, assume the contrary. Then multiplying and rearranging $\beta sr^{\beta}-rs^{\beta}+(1-\beta)r^{\beta+1}\leq 0$. Dividing by r^{β} and regrouping gives $\beta s+(1-\beta)r\leq s^{\beta}r^{1-\beta}$. But the left-hand side is the weighted arithmetic mean of s and r, and the right-hand side is the geometric mean. It is a well-known mathematical theorem (see G. H. Hardy, J. E. Littlewood, and G. Pólya, p. 17) that if $s\neq r$, the arithmetic mean must exceed the geometric mean. Therefore the inequality is false for $s>r$, and $N^*>1$. Regardless of price elasticity E_0, vertical integration for Cobb-Douglas production function increases final product price, a conclusion reached by Schmalensee, Hay, and Warren-Boulton for isoelastic demand curves of final product.

D. $\sigma^*>E_0$

Allen (p. 373) proves that $\sigma^*>E_0$ implies positive cross elasticity $d\ ln\ x/d\ ln\ s$ of *mutatis mutandis* derived demand. Here I show that, with $E_0>1$, it also implies that marginal revenue exceeds marginal cost of y when upstream monopoly charges the benchmark price s^*. That is $s_1<s^*$. Vertical integration must raise product price.

The proof is simple. From (22)

$$N^*=[(s^*-r)/s^*]\times[c^*/(c^*-c_0)][\beta^*+\alpha^*\sigma^*/E_0]$$

The last bracket is the weighted mean of 1 and σ^*/E_0, here obviously greater than 1. The product of the first two brackets is not less than one because *MCF* is concave and its c-intercept is nonnegative.[13] Thus $N^*>1$, and $p_0>p_1$.

E. $\sigma^*=0$

The isoquant evaluated at benchmark price s^*, but not necessarily for $s<s^*$, is in this case parallel to the x-axis. That is, $b^*=b(s^*)$ represents a technologically fixed minimum quantity of y required per unit of final output, defined on the ridge line where marginal product of x is zero. In such situations vertical merger *cannot* lead to price increase: $p_0\leq p_1$.

Technological conditions generate linear *MCF* with positive slope at s^*, and output-compensated demand of Figure 2 is horizontal at b^*—completely price-inelastic ($e^*=0$) in the neighborhood of s^*. The second term of (25) vanishes so that

$$N^*=\frac{(s^*-r)c'^*}{c^*-c_0}$$

[12] *MCF* for the Cobb-Douglas production function is $c=A^{-1}\alpha^{-\alpha}\beta^{-\beta}w^{\alpha}s^{\beta}$, $w=1$, where A is total factor productivity. Because of duality, it is Cobb-Douglas in input prices.

[13] Because *MCF* is concave with nonnegative intercept $c^*/s^*\leq c_0/r$ for $s^*\geq r$. Thus $c^*/r-c^*/s^*\geq c^*/r-c_0/r$ and $[(s^*-r)/s^*][c^*/(c^*-c_0)]\geq 1$.

which, as shown in footnote 10, is because of convexity, always equal or less than unity.

One possibility is that input coefficients take on identical values for s^* and r. Then $c'^*=b^*=b_0$ and $c^*-c_0=(s^*-r)b_0$. Substituting, one obtains $N^*=1$. As has been shown by a number of writers, if isoquants are L shaped, with production coefficients fixed, as in the standard Leontief interindustry model (*CES* and $\sigma=0$), then vertical integration has no effect on product price. This is clearly the outcome also if substitution is possible for $s<r$ but impossible for $s\geqslant r$. In either case *MCF* connects (r, c_0) with (s^*, c^*) by a straight line segment, and, in Figure 2, the horizontal marginal and average curves would coincide for $s^*>r$. Regardless of how low E_0 (i.e., how large s^*), the distances *GJ* and *GH* are identical (infinite).

More generally $e^*=0$, but $e>0$ for *some* prices between s^* and r. At price $\hat{s}$ ($s^*>\hat{s}>r$), the isoquant becomes parallel to the x-axis at the ridge line, and y is limitative (see Nicholas Georgescu-Roegen, pp. 366 ff.). The benchmark input-output coefficients then lie on the ridge line, that is, $b(\hat{s})=b(s^*)$ and $a(\hat{s})=a(s^*)$; but $b^*<b_0$ and $a^*>a_0$. Algebra gives $N^*<1$. Graphically, *MCF* is linear for $s^*>\hat{s}$, but not for $s^*>r$, and the marginal curve in Figure 2 for $s^*>\hat{s}$ would be horizontal, while the average falls. So $GJ<GH$. Upstream monopoly, taking advantage of the inability of downstream firms to substitute against y at the margin, charges $s_1>s^*$. Therefore, $p_0<p_1$.

In sum, when $\sigma^*=0$, vertical integration lowers product price, except under the most pessimistic of technological assumptions that $\sigma=0$ over an entire range, from r to s^*, of relative input prices. In that extreme case price (and output) will be unaltered through merger.

F. $0<\sigma^*<E_0$

When elasticity of substitution is in this range, vertical integration may increase or decrease price. The outcome depends on the characteristics of the production function and the magnitudes of σ^* and E_0. Formula (27) shows that if, at the benchmark, y is utilized, i.e., $c'^*>0$, then $N^*\gtreqless 1$ according as $d/ds\; c'^*/[(c^*-c_0)/(s^*-r)]\lesseqgtr 0$. Thus, following integration, product price would fall (rise) if the ratio of marginal to average evaluated at the benchmark in Figure 2 rises (falls).

An interesting corollary follows. Suppose that for a given production and product demand function (input costs r and w are given) $0<\sigma^*<E_0$, and $N^*<1$. Then with production function unchanged, but product demand sufficiently more elastic (but finite), $N^*>1$, always. Geometrically, if the ratio of marginal to average is increasing at the benchmark (the tangents intersect to the right of s^*), then there is a smaller value of s, that would be the benchmark price at higher E_0 for which the ratio of marginal to average is decreasing. This means that the production function that has technical properties which, together with specific value of E_0, cause a decrease in product price following integration, also has the properties to cause an increase in product price when demand for benchmark output is more price elastic.

The proof of this proposition is based on strong concavity of *MCF*, i.e., $\sigma\not\equiv 0$, $s>r$. (i) Necessary for $N^*<1$, as just discussed, is $d/ds(s^*-r)c'^*/(c^*-c_0)>0$: the ratio of marginal to average at s^* increases. (ii) Strong concavity is $(s-r)c'/(c-c_0)<1$ for $s>r$: marginal below average. (iii) The limit $(s-r)c'/(c-c_0)=1$ as $s\rightarrow r$: marginal and average coincide at the start. Clearly (i) and (iii) can hold under continuity only if $d/ds\; (s-r)c'/(c-c_0)<0$, for some s, over the interval $r<s<s^*$. As the ratio, between zero and one, of marginal to average can be viewed a measure of its degree of concavity, a concave *MCF* cannot become less concave ($N^*<1$) at s^*, if it does not become more concave ($N^*>1$) *somewhere* over the interval from r to s^*. The nonincreasing marginal originating at r, at the same ordinate as the average, can rise at s^* relative to the average only if it has first fallen relative to the average at some s between r and s^*.

An important conclusion: vertical integration may raise or lower product price depending on technological details; but for $\sigma^*\neq 0$, price will always rise if the price elasticity of demand is high enough.

For *CES* production functions and $0<\sigma<1<E_0$, I can prove an additional result. If E_o is low enough, vertical integration will always lead to product price decrease. This corrects conclusions of Warren-Boulton and Hay, referred to in the introduction.

To prove the result, first deduce *MCF* for the *CES* production function, itself of *CES* form:

$$c=(1/\gamma)\left[\delta^{\sigma}s^{(1-\sigma)}+(1-\delta)^{\sigma}\right]^{1/1-\sigma}$$

Using usual notation and nomenclature, γ is the "efficiency" parameter and δ, the "distribution" parameter. Substitution in (25) yields

$$N^*=\frac{\left(\frac{c}{c-c_0}\right)\left(\frac{s-r}{s}\right)\delta^{\sigma}s^{(1-\sigma)}+\left(\frac{s-r}{s}\right)\sigma(1-\delta)^{\sigma}}{\delta^{\sigma}s^{(1-\sigma)}+(1-\delta)^{\sigma}}$$

where asterisks are left off for simplicity. For $\sigma<1$, *CES* isoquants do not intersect the axes, so *MCF* is unbounded and N^* exists for all E_0, that is, for all $s^*>r$. Remember that $c^*/(c^*-c_0)=E_0>1$ is lowered and $(s^*-r)/s^*<1$ is raised as s^* is increased. Further, each of these ratios is unity in the limit, as s^* increases without bound. The product of the two ratios, as proven in footnote 13, however, is never less than unity. Thus N^*, here the weighted arithmetic mean of $[c^*/(c^*-c_0)]\;[(s^*-r)/s^*]$ and $[(s^*-r)/s^*]\;\sigma$, approaches, for $\sigma<1$, a value of N^* between σ and unity as s^*, at the benchmark output, is increased, i.e., as E_0 is lowered toward unity. And that completes the proof. Observe that $N^*>1$ for large E_0, provides here an illustration of the proposition, just demonstrated, that $N^*<1$ for some value E_0, requires $N^*>1$ for a higher value of E_0.

REFERENCES

R. G. D. Allen, *Mathematical Analysis for Economists*, London 1938.

K. J. Arrow et al., "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, *63*, 225–50.

Ward S. Bowman, Jr., *Patents and Antitrust Law*, Chicago 1973.

M. L. Burstein, "A Theory of Full-Line Forcing," *Northwestern Univ. Law Rev.*, Feb. 1960, *55*, 62–95.

Nicholas Georgescu-Roegen, *Analytical Economics*, Cambridge, Mass. 1966.

G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge 1952.

G. A. Hay, "An Economic Analysis of Vertical Integration," *Ind. Organization Rev.*, No. 3, 1973, *1*, 188–198.

J. R. Hicks, *The Theory of Wages*, New York 1968.

D. L. Kaserman, "Theories of Vertical Integration: Implications for Antitrust Policy," *Antitrust Bull.*, Fall 1978, *23*, 483–510.

L. McKenzie, "Ideal Output and the Interdependence of Firms," *Econ. J.*, Dec. 1951, *61*, 785–803.

P. A. Samuelson, "Parable and Realism in Capital Theory: The Surrogate Production Function," *Rev. Econ. Studies*, June 1962, *29*, 193–206.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 442–49.

J. M. Vernon and D. A. Graham, "Profitability of Monopolization by Vertical Integration," *J. Polit. Econ.*, July/Aug. 1971, *79*, 924–25.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, July/Aug. 1974, *82*, 783–802.

A Theory of Monopoly Pricing Schemes with Demand Uncertainty

By MILTON HARRIS AND ARTUR RAVIV*

A few types of monopolistic pricing schemes are commonly used in the marketing of many different products. The most commonly used scheme is the simple single price strategy in which a seller posts a price and offers to sell to anyone wishing to purchase at this price. Another widely used marketing technique is some form of auction, for example, Treasury bills and bonds, some corporate financial securities, art objects, oil leases, government contracts, etc. There are several types of auction procedures classified by the way in which bids are solicited (sealed bid or open) and by the method for determining the final allocation as a function of the bids (for example, competitive or "second price" auctions, discriminating auction, etc.). A third scheme is one in which various prices are charged and buyers paying higher prices are assigned higher priority in receiving the product. We call this technique "priority pricing." For example, natural gas and electric power are sold to some industrial consumers using a priority pricing scheme (users paying lower prices are cut off before those paying higher prices in times of shortage). Another example is the close-out sale in which the price is reduced over time and buyers willing to accept lower probabilities of obtaining the product may be able to purchase it at lower prices. A third example is the way in which mail delivery is priced by most post offices; lower-priced third-class mail is handled only after all first-class mail has been serviced.

The purpose of this paper is to derive *endogenously* the form of an optimal marketing scheme in a context in which almost any conceivable mechanism is feasible. We thereby seek to provide an explanation of the use of the *types* of schemes discussed above and to identify the conditions under which each *particular* scheme will be used. For example, when would one expect to observe a single-priced strategy as opposed to some form of auction.[1]

Previous work in this area has proceeded by imposing a *given* pricing scheme and analyzing optimal values for the parameters of this scheme. David Baron (1971), Duncan Holthausen, and Hayne Leland considered the single price scheme and characterized the optimal single price allocation. Holthausen also considered a scheme in which quantity is set prior to the realization of demand with price determined by *ex post* market clearing. Others have extended this type of analysis to allow different prices to be set *ex ante* in different periods depending on expected demand in those periods. This type of pricing scheme, generally referred to as peak load pricing, has been analyzed by Roger Sherman and Michael Visscher, M. A. Crew and P. R. Kleindorfer, Gardner Brown and M. Bruce Johnson, Robert Meyer, and others. Most of these studies characterize optimal peak load prices and capacity, *given* the peak load pricing scheme. Joseph Stiglitz considered the possibility of price discrimination based on quantity purchased, that is, nonlinear pricing. He characterizes optimal payment schedules which give total cost to a consumer as a function of the quantity purchased.

Another pricing scheme analyzed in the literature is similar to what we have called priority pricing. Maurice Marchand and

*Associate professor, Graduate School of Industrial Administration, Carnegie-Mellon University; and associate professor, Graduate School of Industrial Administration and Faculty of Management, Tel Aviv University, respectively. Comments of an anonymous referee and George Borts are gratefully acknowledged along with the support of the National Science Foundation, grant SOC-7826262.

[1] Obviously, pricing schemes other than the ones discussed above are also observed, namely, price discrimination based on directly observable characteristics of the buyers. We do not seek to explain these schemes but instead focus on situations in which the buyers are identical in terms of their directly observable characteristics.

John Tschirhart and Frank Jen consider a scheme in which an interruptible service is priced according to its reliability. Given this pricing scheme, these studies characterize an optimal price-reliability schedule.

A separate literature analyzes auctions as a method of marketing goods. Most studies in this area analyze the properties of *given* auctions (see, for example, Baron, 1972, Charles Holt, Steven Matthews, John Riley and William Samuelson, and Robert Wilson). The object of these studies was generally to compare two different types of auctions based on some criterion such as seller's expected revenue. A somewhat broader view was taken by William Vickrey and Roger Myerson (1981) who searched for an optimal auction design among a larger class of feasible auctions. In our earlier paper, we analyzed optimal auction design, but also proved that in a certain environment an auction is indeed an optimal allocation mechanism.

In the present paper, we provide an explanation of the use of various observed monopoly pricing schemes. Our approach is based on the presumption that the observed market scheme is chosen optimally by the seller from a large class of feasible allocation mechanisms. This class contains the three pricing schemes discussed above (single price, priority pricing, and auctions) as well as most other conceivable schemes (for example, non-linear pricing). Thus, in contrast tc previous studies, we do not impose a particular marketing technique, but instead derive an optimal scheme *endogenously*. This deeper approach allows us to explain observed marketing schemes.

Our model consists of a single, monopolistic seller and N potential buyers. The seller produces a homogeneous product with constant marginal production cost up to a capacity limit. The capacity limit may or may not be binding. Each buyer is assumed to demand up to one unit of the product at any price at or below his reservation price. Buyers are identical except for reservation price. A central assumption of the model is that there is asymmetric information among the agents. In particular, each buyer knows only his own reservation price and not that of any other buyer. The seller does not observe any buyer's reservation price. Each agent has Bayesian priors regarding reservation prices he cannot observe. Our approach to deriving optimal marketing schemes is based on the methodology suggested by Harris and Robert Townsend, and is described briefly in Section II.

Our results consist of characterizing optimal marketing schemes for the above environment. It turns out that the optimal marketing technique depends crucially on the assumptions regarding the capacity limit. In particular, we show that

(i) When potential demand exceeds capacity, the priority pricing scheme discussed above is optimal for the seller. We show explicitly how the optimal priority prices depend on the exogenous aspects of the model, namely, marginal cost, capacity, and the priors of the agents. Another scheme which is optimal in some environments is a modified version of the (Vickrey) competitive auction. In this auction, buyers may submit sealed bids *above a minimum acceptable bid* and a uniform price is charged to all accepted bids. The minimum acceptable bid is equal to the lowest priority price. Again we show how the equilibrium price depends on the bids and the exogenous aspects of the model.

(ii) When capacity exceeds potential demand, the single price scheme is optimal. The optimal price is shown to be determined by the usual "marginal revenue equals marginal cost" condition modified to account for the specific type of uncertain demand assumed in the model.

(iii) When capacity can be chosen by the seller, for sufficiently low cost of capacity, it is optimal to choose capacity equal to maximum potential demand and charge a single price.

The explanation of the use of various pricing schemes provided by these results appears to be consistent with the observations described at the beginning of this section. In particular, we seem to observe priority pricing or auctions mainly when capacity limitations are important. On the other hand, we observe a single monopoly price mainly when capacity constraints are not binding. Obviously oil leases, securities, rare art objects, government contracts are sold by monopolists and potential demand for these products exceeds capacity. As mentioned above, these products are generally sold at auction as predicted by our analysis. Similarly discontinued products or styles, natural gas, electric power, mail delivery, are also examples of products in which capacity is limited relative to potential demand. These products are often sold by priority pricing. Furthermore most products sold by monopolists at a single price are products for which capacity limitations are unimportant or increases in capacity are relatively cheap. Other, more specific positive implications of the analysis are discussed below:

I. The Economic Model

A. *Production Technology and Consumers' Preferences*

We consider a simple "partial-equilibrium" environment populated by a monopolistic seller and N potential customers. The monopolist is described by his marginal cost function, his preferences, and the information he possesses regarding demand for his product. In order to keep the analysis as simple as possible, we suppose that marginal cost is a constant, c, for output below some capacity limit, Q (see Figure 1).[2] It is not possible to produce more than Q units of output. Initially (in Section IV) we assume that the capacity limit is exogenously or historically fixed. Later (in Section V) we relax this assumption to consider the case in which capacity can be chosen by the monopolist. We also assume that output can be chosen after acquiring information about demand. This information is acquired through the mechanism used to market the product. The objective of the monopolist is assumed to be expected profit maximization. Because the monopolist does not have perfect information about the demand for his product until after choosing a pricing scheme and actually selling the product, he must consider profits to be random when he makes his decisions. The exact nature of this randomness will be discussed below in connection with the description of the monopolist's information about demand.

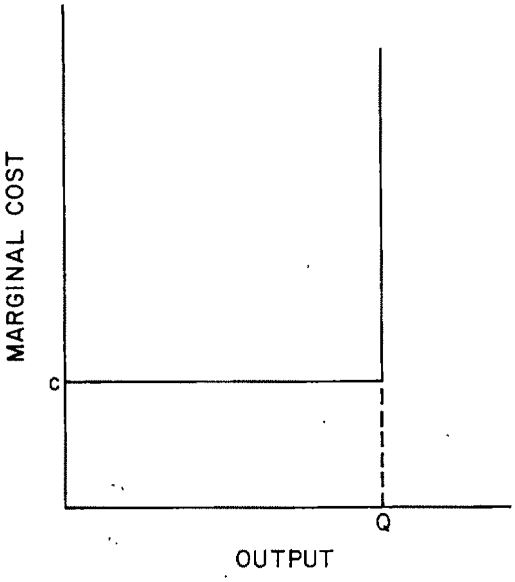


FIGURE 1. THE MONOPOLIST'S COST STRUCTURE

Each of the N potential customers in our model is characterized by his marginal willingness to pay for the monopolist's product, his endowment of money, and his information concerning the demands of other buyers and the cost structure of the monopolist. Again in order to simplify the analysis, we assume that a typical buyer's demand for the product is given by a unitary demand curve. That is, each potential customer $i(i=1,\dots,N)$ has a fixed reservation price (or marginal willingness to pay) denoted R_i, and is willing to buy at most one unit of the product at any price $p \leqslant R_i$ and no units at prices above R_i (see Figure 2).[3] Note that the reservation prices may differ across customers. In what follows, it will be con-

[2] This assumption has been used extensively in the literature on imperfect competition under demand uncertainty (see, for example, Brown and Johnson, Crew and Kleindorfer, and Meyer).

[3] This assumption is standard in the auctions literature (see, for example, Vickrey, Wilson, Matthews). It has recently come to our attention that Eric Maskin and Riley have been working on an extension of the present model to the case of linear downward-sloping demand.

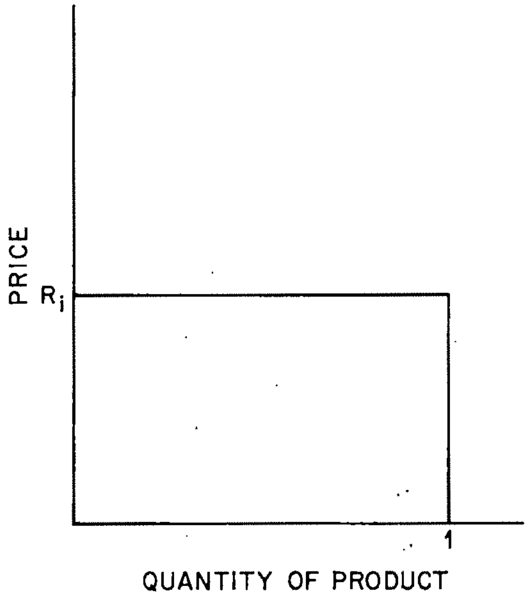


FIGURE 2. DEMAND CURVE OF A TYPICAL CUSTOMER

venient to have a representation of the individual consumer demands in the form of utility functions. This utility for buyer i, $U(d_i, q_i, R_i)$, is a function of the amount of money he holds, d_i, the quantity of the good he consumes, q_i, and his reservation price R_i. The form of the function U which corresponds to the unitary demand function described above is

$$(1) \quad U(d_i, q_i, R_i) = d_i + R_i \min(q_i, 1)$$

Thus, for $q_i < 1$, buyer i has a fixed marginal rate of substitution R_i between the good and money, that is, his marginal willingness to pay for the good R_i is independent of the quantity he consumes for quantities less than 1. Moreover, buyer i has no utility for amounts of the good in excess of one unit, that is, his marginal willingness to pay for quantities in excess of one unit is zero. In all respects other than reservation price, the buyers are assumed to be identical. In particular, each customer starts off with $\bar{d}$ units of money (dollars) to spend on this good.

B. *The Information Structure: Who Knows What and When*

To complete the description of the model, we must specify the information available to each agent. If the seller knew each consumer's true reservation price with certainty, it is clear that his optimal strategy would be simply to sell one unit to each of the Q highest reservation price buyers and charge each such buyer his true reservation price. Thus a crucial assumption of the model is that the monopolist cannot tell buyers apart. That is, the monopolist does not know the true reservation price of any consumer unless that price is revealed through the consumer's purchasing behavior or voluntarily revealed by the consumer. In order to model the monopolist's imperfect information regarding the reservation prices of the buyers, we assume that the monopolist has some prior beliefs about these prices. The simplest such assumption is that the monopolist views each R_i as having equal probability of being any one of the numbers $x_1, \ldots, x_k$. That is, in making his decisions about how to price his product, the monopolist acts as if the R_i's are drawn at random independently from a uniform distribution on the set $X = \{x_1, \ldots, x_k\}$. (All of the major results would continue to hold with a nonuniform distribution on X.) It is in this sense that the demand faced by the monopolist is random, that is, no individual consumer's demand is subject to random shocks, but the monopolist, because of his ignorance, is forced to view demand as uncertain.[4] Other than the actual reservation prices of the buyers, the monopolist knows everything else about the environment, namely his marginal cost c, his capacity Q, the number of potential buyers N, and the monetary resources of each buyer $\bar{d}$.

Just as the monopolist is assumed not to know the true reservation prices of the

[4] Obviously, we are taking a Bayesian view of decision making under uncertainty. We believe this view to be appropriate. Again, it is standard in the auctions literature. Note that because of the monopolist's uncertainty regarding individual demand, aggregate demand is also uncertain from his point of view, unless the number of buyers N is very large.

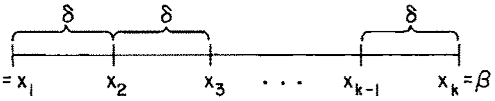


FIGURE 3. THE SET OF POSSIBLE RESERVATION PRICES, X

buyers, each buyer is assumed to be informed only about *his own* reservation price. Thus buyer i knows R_i but doesn't know R_j for some other buyer j. Again we model buyer i's uncertainty about R_j by assuming that buyer i believes that R_j has an equal chance of being any of the numbers $x_1,\ldots,x_k$. Finally, each buyer is informed about all other aspects of the environment, namely, c, Q, N, and $\bar{d}$.

Finally, there are two more assumptions which will simplify the analysis but will not affect the qualitative results regarding when it is optimal to use which kind of pricing mechanism. The first of these is that the possible reservation prices x_i are equally spaced on an interval $[\alpha,\beta]$ as in Figure 3. The distance between any two successive points is denoted $\delta>0$. Later (in Section V) we will investigate the results as $k\rightarrow\infty$ and $\delta\rightarrow 0$, i.e., as X approaches the interval $[\alpha,\beta]$. Second, we assume that each buyer has enough money to purchase one unit, even at the highest reservation price, i.e., $\bar{d}\geqslant x_k$.

The remainder of the paper is devoted to characterizing marketing schemes which are optimal for the seller. For example, would we expect the monopolist to sell his product by setting a monopolistic price on a take-it-or-leave-it basis (i.e., the approach usually assumed in textbooks), or by adopting some more complicated scheme in order to discriminate among buyers? Obviously any attempt at discrimination is limited by the fact that the seller does not know the reservation prices of individual buyers. Nevertheless, some discrimination is possible; for example, using auctions. We will see below that under certain conditions, simple monopoly pricing is optimal, whereas under other conditions, more complicated pricing schemes prove superior.

II. The Monopolist's Problem

Having set out the model, we must now specify exactly the problem which the monopolist solves in order to discover an optimal scheme or mechanism for pricing and selling his product. In formulating this problem, we imagine that the monopolist has at his disposal a large class of alternative pricing schemes. This class of schemes includes both simple one-stage mechanisms as well as more complicated, sequential ones. The monopolist then searches over this class of schemes until he finds one which results in maximal expected profits.

In order to make our optimality results as strong as possible, we include in the class of available pricing schemes almost any mechanism which one can imagine for pricing and selling a homogeneous product.[5] Perhaps the simplest imaginable scheme is to announce a price and sell Q units to the first Q buyers who are willing to pay that price. We call this the "single price" scheme, and we shall see that in some cases, this scheme is optimal for the monopolist. Another simple type of scheme is a sealed-bid auction. In this mechanism, each potential buyer is asked to write down a bid for one unit, that is, to write down a number which is interpreted as his declared reservation price. After all buyers have submitted their bids, the number of units allocated to each buyer and the price each must pay is determined following some pre-announced allocation rule as a function of the submitted bids. Various types of sealed

[5] Formally, we define a *mechanism* as a multistage (or sequential) game which consists of two objects. First is a finite sequence of sets (or message spaces), one for each player, which define the set of feasible plays or messages for each player at each stage. The set of feasible plays for an agent at stage s could depend on the history of the game up through stage $s-1$. Second is a function, called the *allocation rule* of the mechanism, which determines the final allocation as a function of the entire history of the game (i.e., as a function of the sequence of actual plays or messages). A mechanism is *feasible* if it results in a feasible allocation of goods and money for any sequence of plays by the agents. The class over which the monopolist searches is simply the set of all feasible mechanisms. These definitions are adapted from Harris and Townsend. A more precise and formal treatment may be found there.

bid auctions can be represented by different allocation rules. For example, the U.S. Treasury often uses a discriminating sealed bid auction for pricing and selling Treasury bills. In this auction, the Q highest bidders each receive a unit and each pays his bid price. This is an example of a nonsequential scheme since all buyers submit bids simultaneously (or equivalently, no bid is announced until all bids are in). An example of a sequential scheme is the "open English" auction. In this scheme, buyers announce bids openly. After a bid is announced, any other buyer may raise the bid. When no one is willing to raise a bid, the highest bidder gets the unit and pays his bid price. All of the above schemes are included in the class of mechanisms which are available to the monopolist. In addition any other scheme, sequential or nonsequential, as long as it always results in a feasible allocation of goods and money (no matter how the buyers behave during the scheme), is allowed.

The problem of the monopolist as described thus far appears to be horribly complicated. First of all, the class of objects over which he must search consists of possibly highly complex schemes (or games). Second, the monopolist, in order to evaluate a given scheme, must take into account the strategic behavior of the buyers who participate in the scheme. That is, the monopolist must take into account the fact that, given a scheme, each buyer will act in his own interests given his expectations about the behavior of other buyers. In view of this complexity, the problem seems completely unapproachable. Fortunately, however, both of the above complications can be vastly simplified by appealing to the following result.

THEOREM 1: *Revelation Principle.*[6] *For any scheme there is an equivalent scheme which is both direct and truthful.*

In the theorem, we use the terms direct and truthful schemes which we must now define. A *direct scheme* is a very simple scheme in which each buyer writes down a bid or *declared* reservation price, r_i (without knowing what the other buyers are writing). The declared reservation price must be one of the possible reservation prices, that is, a member of the set X. Once these declarations have been submitted, an *allocation rule* gives the actual allocation of money d_i, and product q_i, as a function of the declarations. The allocation rule for such a direct mechanism is denoted by

$$[d_i(r), q_i(r)] \qquad i=1,\dots,N$$

where $r=(r_1,\dots,r_N)$. The seller is assumed to produce the total quantity allocated to buyers, $\Sigma_{i=1}^N q_i(r)$, and to receive payments from the buyers totaling $\Sigma_{i=1}^N[\bar{d}-d_i(r)]$. Different types of direct schemes can be modeled by different allocation rules, that is, different d_i's and q_i's. A *truthful scheme* is a direct scheme in which the optimal strategy for any buyer i is to declare his true reservation price, that is, $r_i=R_i$, given that all other buyers are declaring their true reservation prices.[7,8]

[6] It is well beyond the scope of this paper to prove Theorem 1. This theorem is essentially a combination of Theorems 1 and 2 of Harris and Townsend. A result similar to our Theorem 1 may be found in Myerson (1979). The terminology "Revelation Principle" is also due to Myerson.

[7] In the language of game theory, a truthful scheme is a direct game in which truth telling is a Nash (or, more precisely, Bayes) equilibrium. Not all direct schemes are truthful. A somewhat ridiculous example is the direct scheme in which the Q units are allocated to the Q *lowest* bidders at their bid prices. In this scheme, each buyer would declare the lowest possible reservation price x_1 regardless of his true reservation price.

[8] Some rough intuition for Theorem 1 may be gained as follows. First suppose we have a given pricing scheme which is not direct. Suppose also that the equilibrium strategies of the buyers in this scheme as functions of the true reservation prices are known. We can construct an equivalent truthful scheme as follows. Instead of having each buyer compute and report his equilibrium strategy (as a function of his true reservation price), we simply ask each buyer to report his reservation price while informing him that this reported value will be used to calculate the exact same strategies he would have chosen himself. That is, the reported reservation price is plugged into the buyer's equilibrium strategy function. The resulting strategies will then be used to determine the allocation just as in the original procedure. Since the buyers had no reason to lie to themselves about their respective reservation prices when they were

In view of Theorem 1, the search for a profit-maximizing pricing scheme may be confined to direct pricing schemes *without any loss of generality*. Thus, we can ignore complicated sequential schemes such as the open English auction, without any fear that in so doing we have thrown out schemes which might result in higher expected profits. Moreover, recall that all direct schemes are very similar. They all require the buyers simultaneously to submit declared reservation prices. Direct schemes differ only with respect to their allocation rules, that is, with respect to how the goods and money are allocated as a function of the declared reservation prices. Therefore, in order to search over direct schemes, we need only search over feasible allocation rules $[d_i(r), q_i(r)]$.

We can, however, using Theorem 1 simplify the problem still further. Namely, we can, also *without loss of generality*, restrict attention to truthful schemes. Now, if we are going to restrict our attention to truthful schemes, we must have some way of telling these from other direct schemes simply by inspecting allocation rules. Recall that in a truthful scheme, each buyer's expected utility is higher if he declares his true reservation price than if he declares any other possible reservation price, given that everyone else is also telling the truth. Therefore, the allocation rules of truthful schemes, and only those allocation rules, satisfy the following *self-selection* (SS) conditions: for any buyer i and any two possible reservation prices x, y in X,

$$(SS) \quad E_i U[d_i(x, R_{-i}), q_i(x, R_{-i}), x] \geq E_i U[d_i(y, R_{-i}), q_i(y, R_{-i}), x]$$

This condition is simply a mathematical version of the statement that if buyer i's true reservation price is x, then he is better off to declare x (as opposed to any other declaration y) given that all other buyers are declaring truthfully. In the mathematical expression, the symbol R_{-i} refers to the vector of true reservation prices of buyers other than i. The expression $d_i(x, R_{-i})$ means $d_i(R_1,\ldots, R_{i-1}, x, R_{i+1},\ldots, R_N)$, and the expressions $q_i(x, R_{-i})$, $d_i(y, R_{-i})$, $q_i(y, R_{-i})$ are to be interpreted similarly. The expectation E_i refers to the expectation over the unknown (to buyer i) reservation prices of the buyers other than i using the prior beliefs described in Section II. The utility function U is the one given in equation (1) of Section I. In order to consider only truthful schemes, we will consider only allocation rules which satisfy (SS).

In addition to (SS), there are several other restrictions which we must place on the allocation rules which the monopolist may choose. First of all, we do not allow the monopolist to force buyers to participate in any pricing scheme. In order to be eligible, a truthful scheme must provide each buyer with at least as much expected utility as he would obtain in autarky (i.e., by retaining his endowment of $\bar{d}$ dollars and not acquiring any goods).[9] The mathematical version of this condition, which we refer to as *individual rationality* (IR) is

$$(IR) \quad E_i U[d_i(R), q_i(R), R_i] \geq \bar{d}$$

for every i and R_i.

Finally, we require that the allocation rule be feasible in the sense that no more than $N\bar{d}$ dollars and no more than Q units of the good may be allocated to the buyers and the allocations may not be negative. These conditions must be fulfilled no matter what the true reservation prices are. Mathematically, we require

$$(T.1) \quad \sum_{i=1}^{N} d_i(R) \leq N\bar{d}$$

computing their own strategies, they will have no incentive in the new scheme to lie about their reservation prices. Consequently, the new scheme will result in exactly the same equilibrium outcome as the original one, and moreover, in the new scheme the equilibrium declarations will be truthful.

[9] This does not prevent the seller from excluding any particular buyer, say i. He can still exclude buyer i simply by setting $d_i(R) \equiv \bar{d}$, $q_i(R) \equiv 0$. If this restriction were not present, it is clear that an optimal scheme would be simply to require each buyer to hand over his $\bar{d}$ dollars without receiving anything in return.

for every R,

$$(T.2) \qquad \sum_{i=1}^{N} q_i(R) \leqslant Q$$

for every R,

$$(T.3) \qquad d_i(R) \geqslant 0, q_i(R) \geqslant 0$$

for every i and R.

The problem which faces the monopolist can now be stated as a simple mathematical programming problem, namely to choose an allocation rule $[d_i(R), q_i(R)]$ which maximizes expected profits subject to the conditions (SS), (IR), and (T).[10] In mathematical form, we assume the monopolist solves

$$[P] \qquad \max_{[d_i(R), q_i(R)]} E \sum_{i=1}^{N} \left[\bar{d} - d_i(R) - cq_i(R)\right]$$

subject to (SS), (IR), and (T). Here E refers to the expectation over R.

The remainder of the paper is devoted to characterizing solutions to problem $[P]$ under various assumptions regarding capacity Q, and its relation to the number of potential buyers N.

III. A Method for Solving the Problem

Our general approach to finding an optimal pricing scheme, that is, solving problem $[P]$, consists of two stages. First, we show that $[P]$ is equivalent to a much simpler problem. Second, we ignore some of the contraints of $[P]$, solve the reduced problem $[P']$, and then show that the solution of $[P']$ satisfies the ignored constraints and therefore solves $[P]$.

Showing that $[P]$ can be simplified without loss of generality involves three steps. First recall from equation (1) that the expected utility appearing in the (SS) and (IR) constraints of problem $[P]$ involves a non-linear expression containing $\min(q_i, 1)$. It is easy to see that any solution of $[P]$ will have $q_i(R) \leqslant 1$ since production is costly and $q_i(R) > 1$ results in the same value of the expected utility expressions of the buyers as $q_i(R) = 1$. Therefore we can replace the utility function in problem $[P]$ by the linear expression $d_i(R) + R_i q_i(R)$ and add the additional constraints $q_i(R) \leqslant 1$, for all i and R.

The second simplification is achieved by noticing that, since all agents are the same except for reservation price, it should be possible to find a solution to $[P]$ in which any two agents who draw the same reservation price get the same bundle. Although their common bundle depends on the other agents' reservation prices, it should only depend on the number of agents having each possible reservation price (i.e., each x_j) and not explicitly on which agents have which reservation prices. More formally, we state

LEMMA: *Problem* $[P]$ *has a symmetric solution, that is, there is a solution* $[d_i^*, q_i^*]$ *of* $[P]$ *such that* $d_i^*(R_1, \ldots, R_i, \ldots, R_N) = d_1^*(R')$, $q_i^*(R_1, \ldots, R_i, \ldots, R_N) = q_1^*(R')$, *for any* $i = 1, \ldots, N$ *and any rearrangement* R' *of the elements of* R *in which* R_i *appears in the first place.*

PROOF:

First we simply note that since $[P]$ is a linear program any convex combination of solutions of $[P]$ is itself a solution. Next, given any solution, we can construct $N! - 1$ other solutions which are the same as the first solution except that the agents are renamed and elements of R are rearranged. Finally these solutions may be averaged to obtain a new solution with the symmetry property claimed.

From now on we will only be interested in symmetric solutions, that is, those that satisfy the conclusion of the Lemma. Now from the Lemma, it follows that

$$(2a) \qquad E_i d_i(x, R_{-i}) = E_1 d_1(x, R_{-1})$$

$$(2b) \qquad E_i q_i(x, R_{-i}) = E_1 q_1(x, R_{-1})$$

[10] Note that we now write the allocation rule as a function of the true rather than the declared reservation prices. This is allowed since we are restricting attention to truthful schemes so that declared and actual reservation prices are identical.

for any $x \in X$. Using these equalities, it follows that the (SS) and (IR) constraints for buyers other than buyer 1 are redundant.

The third simplification involves showing that the (IR) constraints are redundant except for $j=1$. The proof consists of rewriting (SS) and (IR) using the above results, then noting that since $x_j > x_{j-1}$, it follows from (SS) that

$$E_1\left[d_1(x_j, R_{-1}) + x_j q_1(x_j, R_{-1})\right]$$

$$\geqslant E_1\left[d_1(x_{j-1}, R_{-1}) + x_{j-1} q_1(x_{j-1}, R_{-1})\right]$$

Thus if (IR) holds for $j=1$, it holds for all $j>1$ as well. Therefore we can, without loss of generality, drop all (IR) constraints except the one for $j=1$.

Finally, by summing equations (2a) and (2b) over $i=1,\dots,N$ and taking expectations with respect to x, we obtain

$$E \sum_{i=1}^{N} d_i(R) = NEd_1(R),$$

$$E \sum_{i=1}^{N} q_i(R) = NEq_1(R)$$

Consequently, using all of the above results, problem $[P]$ is equivalent to

$$[P] \quad \max_{[d_i(R),\, q_i(R)]} E\pi$$

$$= NE\left[\bar{d} - d_1(R) - cq_1(R)\right]$$

subject to

$$(SS) \quad E_1\left[d_1(x_j, R_{-1}) + x_j q_1(x_j, R_{-1})\right]$$

$$\geqslant E_1\left[d_1(x_m, R_{-1}) + x_j q_1(x_m, R_{-1})\right]$$

$$\text{for } j \neq m,\ j, m = 1,\dots,k$$

$$(IR) \quad E_1\left[d_1(x_1, R_{-1}) + x_1 q_1(x_1, R_{-1})\right] \geqslant \bar{d}$$

$$(T.1) \qquad \sum_{i=1}^{N} d_i(R) \leqslant N\bar{d} \qquad \text{for any } R,$$

$$(T.2) \qquad \sum_{i=1}^{N} q_i(R) \leqslant Q \qquad \text{for any } R,$$

$$(T.3) \qquad d_i(R) \geqslant 0,\ 1 \geqslant q_i(R) \geqslant 0$$

for any i and R, and $[d_i, q_i]$ satisfy the symmetry conditions of the Lemma.

This completes the derivation of a simpler problem which is equivalent to $[P]$. Hereafter $[P]$ refers to this simpler equivalent problem. We do not solve $[P]$ directly, however. Instead it is convenient to ignore some of the constraints of $[P]$. This leads to an even simpler problem with a larger constraint set, which we call problem $[P']$. If we can solve $[P']$ and verify that the solution also satisfies the dropped constraints of $[P]$, then we will have found a solution of $[P]$. The constraints which are ignored are all the (SS) constraints in which $m \neq j-1$ and the constraints $d_i(R) \geqslant 0$. It is easy to show that any optimal solution of $[P']$ will satisfy all the (remaining) (SS) constraints and the (IR) constraint as equalities. These equalities and $x_j - x_{j-1} = \delta$ can then be used to derive the following

$$(3) \quad E_1 d_1(x_j, R_{-1})$$

$$= \bar{d} + \delta \sum_{m=1}^{j-1} z_m - z_j\left[x_1 + (j-1)\delta\right]$$

where

$$(4) \quad z_j = E_1 q_1(x_j, R_{-1}) \qquad \text{for } j = 1,\dots,k$$

Substituting (3) into the objective function yields, after some manipulation

$$(5) \quad E\pi = N\varphi \sum_{j=1}^{k} \left[x_1 + (2j-k-1)\delta - c\right] z_j$$

where $\varphi = 1/k$. In order to solve $[P']$, we maximize the right-hand side of (5) with respect to the z_j's subject to the constraints implied by problem $[P']$. To do this, we must first introduce the concept of an allocation by rank. Suppose we have a particular ranking of the buyers (the ranking could involve ties, i.e., more than one buyer could receive the same rank). An *allocation by rank* is the allocation in which units of the product are assigned to buyers according to their rank, from highest to lowest. That is, each buyer with the highest numbered rank is allocated one unit (or his proportional share of the Q units if more than Q buyers have the highest rank). The remaining units are then allocated in the same way to buyers of the second highest rank, etc. This continues until all Q units are allocated. An *allocation by rank with cut-off rank* ρ is an allocation by rank with the additional condition that buyers with rank strictly below ρ are not assigned any units even if less than Q units are assigned to buyers with rank ρ and above.

In order to maximize the right-hand side of (5), note that the expression is linear in the z_j's and the coefficients of the z_j's increase with j. Therefore let T be the smallest j for which the coefficient of z_j is nonnegative, that is, T is such that

$$(6)\quad \begin{aligned} x_1+\delta(2j-k-1)&\geqslant c \quad \text{for } j\geqslant T\\ &< c. \quad \text{for } j<T \end{aligned}$$

Consequently to maximize the right-hand side of (5) we choose $z_1, z_2,\ldots, z_{T-1}$ as small as possible, then choose z_k as large as possible, z_{k-1} as large as possible given our choice of z_k, etc., down to z_T. Because of the nonnegativity constraints, we must have $z_j\geqslant 0$ for each j. Therefore, we choose $z_1^*=z_2^*=\ldots=z_{T-1}^*=0$. To make z_k as large as possible, one should choose $q_1(x_k, R_{-1})=1$ for all R_{-1} (since $q_i(R)\leqslant 1$ for all R), that is, give buyer 1 one unit when his reservation price is the highest one possible. Because of the symmetry constraint, however, this requires giving any buyer whose reservation price is x_k one unit. If $Q<N$, the number of buyers with reservation price x_k might exceed Q. In this case, the largest z_k is obtained by sharing the Q units among all buyers with reservation price z_k. Once z_k^* is determined, the maximization of z_{k-1} is achieved by allocating the remaining units in the same fashion. This continues until z_T^* is determined. It should be clear from this discussion that maximizing the right-hand side of (5) is achieved by choosing $q_1(x_j, R_{-1})$ to be the quantities which are assigned to agent 1 in an allocation by rank with cut-off rank T where the ranking of agents is by true reservation price, that is, assign each agent with reservation price x_j rank j. The maximizing values $z_T^*,\ldots, z_k^*$ are then the values which results from choosing $q_1(x_j, R_{-1})$ in this way. Explicit expressions for the values z_j^* are needed only in the formal proofs and are therefore given only in the Appendix.

Most of the work of solving $[P']$ has now been done. The maximizing choices $z_T^*,\ldots, z_k^*$, however, depend on whether $Q<N$ or $Q\geqslant N$. Since N is the maximum possible demand, we refer to these two cases as "potential demand exceeds capacity ($Q<N$)" and "capacity exceeds potential demand ($Q\geqslant N$)". We analyze these two cases separately in the following section.

IV. Optimal Pricing Schemes with Exogenous Capacity

In this section we assume that the capacity of the seller, Q, is exogenously fixed. The analysis of optimal pricing schemes under this assumption is important for two reasons. First, in many observed situations this assumption does, in fact, characterize the situation. For example, the quantity of Rembrandt originals is historically fixed. Other examples of this type come easily to mind (rare coins, stamps, etc.). Another example is the fixed capacity of a given airline flight. Also many manufacturing facilities are characterized by fixed capacity, at least in the short run (utilities, etc.). Another example occurs at the end of a season or model year when no further production takes place and existing stocks are simply disposed of. As will be seen, our analysis explains the procedures often observed in these fixed-capacity environments. Second, the analysis of the fixed-capacity case is a necessary first step in examining the case of endogenous capacity which we pursue in Section V.

A. *Potential Demand Exceeds Capacity* ($Q<N$)

Clearly, if the seller had full information about the reservation price of each buyer, it would be optimal for him to sell one unit each to buyers with the highest reservation prices and to charge each buyer his true reservation price, that is, to discriminate perfectly among buyers. In the absence of such full information, it is not possible to achieve such perfect discrimination. It is possible, however, to discriminate to some extent by charging some buyers higher prices in exchange for higher priority access to the product. Higher reservation-price buyers will be willing to pay somewhat higher prices, relative to low reservation-price buyers, in exchange for higher priority. This fact can be used by the seller as a basis for discrimination. Our results for this subsection exhibit two optimal pricing schemes each of which exploits this idea. One of these two schemes is optimal for any values of the parameters of the model (i.e., N, Q, k, $\bar{d}$, etc.). The other is optimal only for some values of these parameters.

The first scheme we analyze is a "priority pricing scheme." A *priority pricing scheme* is one in which the seller announces a schedule $p_m < \dots < p_n$ of priority prices. Each buyer chooses a priority price which he is willing to pay. Buyers are then ranked by the priority prices they choose, that is, buyers choosing higher priority prices are assigned higher rank. Buyers not wishing to buy at any of these prices simply do not announce a choice of price and are assigned rank $0 < m$. The product is then allocated by this ranking with cut-off rank m. Each buyer pays his chosen priority price times the quantity he receives. We show in the next theorem that a priority pricing scheme is an optimal mechanism provided that the priority prices are properly chosen. A sketch of the proof of this theorem, and all subsequent theorems, is given in the Appendix.

THEOREM 2:

A. *When $Q < N$, an optimal scheme is the priority pricing scheme with priority prices $p_T^*, \dots, p_k^*$ where T, the cut-off rank, is defined by* (6), *k is the number of possible reservation prices, and*

$$(7) \qquad p_i^* = x_i - (\delta/z_i^*) \sum_{j=T}^{i-1} z_j^*,$$

$$\textit{for } i = T+1, \dots, k, \; p_T^* = x_T$$

($z_T^*, \dots, z_k^*$ *are the numbers defined above*).

B. *Any buyer with true reservation price x_i chooses priority price p_i^* when faced with the above scheme, for $i = T, \dots, k$. Buyers with reservation price less than x_T will choose not to buy.*

By adopting the scheme described in Theorem 2, the seller in effect converts a homogeneous product into a heterogenous one. This is accomplished by attaching to the product various probabilities of being allowed to purchase it according to the priority scheme. The seller is thus enabled to discriminate across buyers: buyers with higher reservation prices will choose to pay higher priority prices in exchange for higher reliability (probability of obtaining the product). Notice, however, that the seller cannot capture all the rents possible under full information since each priority price p_i^* is strictly less than its corresponding reservation price x_i, except for the lowest price p_T^* (see equation (7)). This loss of rents reflects the fact that the seller cannot directly observe buyer's preferences.[11]

Under the optimal scheme of Theorem 2, it may happen that not all Q units will be sold. This occurs when fewer than Q buyers have reservation prices at or above $p_T^* = x_T$. In this case, the seller could have sold more units if he were willing to offer priority prices below p_T^*. Moreover, there could be prices below p_T^* which are still above marginal cost (see equation (6)). The question then arises why not sell more units at these lower prices. Suppose that the seller decided to reduce the lowest priority price in order to sell more units. Obviously, in order to attract any additional buyers, the lowest price must be reduced at least to x_{T-1} (from x_T). In this case, however, buyers with reservation price x_T would find it advantageous to choose priority price x_{T-1} instead of x_T. Similarly buyers with reservation prices above x_T would find it in their interest to choose lower priority prices than before. The consequent loss of revenue more than compensates for the reve-

[11] The reader might wonder why it is not optimal to charge priority prices $x_T, \dots, x_k$ instead of $p_T^*, \dots, p_k^*$ since for $i > T$, $x_i > p_i^*$. The reason is that if the prices were $x_T, \dots, x_k$, any buyer with reservation price x_i would *not* choose to pay priority price x_i but would choose to pay price x_j where x_j is the reservation price just *below* p_i^*.

nue gained from any additional sales. This is why it may be optimal to set the lowest priority price such that for some realizations of demand, less than Q units are sold.

The priority pricing scheme shown to be optimal in Theorem 2 is in fact often observed in reality. For example, end of model year or end of season sales can be viewed as priority pricing schemes. In such sales, prices are gradually reduced and some buyers pay lower prices than others while taking greater risk of not being able to purchase the product. Our model predicts that in these sales, the schedule of price reductions will be independent of remaining quantities. This appears to provide a means of testing the model. Another example of priority pricing occurs in the provision of natural gas. Here some industrial users of natural gas pay higher prices in exchange for being last to be cut off in case of shortage. Standby tickets at lower prices offered by airlines provide another example. A similar example involves the sale of tickets for Broadway plays in New York: at 4:00 P.M. remaining tickets (if any) for that evening's performance go on sale at half price.

Another scheme which may also be optimal is a common price auction with minimum acceptable bid. A *common price auction* in this environment is a sealed bid auction in which each buyer submits a bid for one unit of the product. A bid can be *any* reservation price, that is, any element of X. The bids determine the common price paid by every "successful" bidder and the number of units allocated to each bid. Different common price auctions are characterized by different methods of determining this price as a function of the bids. A *common price auction with minimum acceptable bid b* is a common price auction with the additional feature that any bid below b is rejected automatically regardless of whether or not there are Q bids at or above b.

It turns out that for some values of the parameters N, Q, and $\bar{d}$, one can construct a common price auction which yields the same expected profit as the priority pricing scheme just discussed. In this auction, the seller allocates the Q units by rank where the rank corresponds to the bid. The seller also sets a minimum acceptable bid given by x_T, that is, the cut-off rank is T as in the priority pricing scheme (see equation (6)). The common price paid by all successful bidders is determined as follows. Suppose that the lowest bid at which any units are awarded in the allocation by rank (the lowest accepted bid) is x_i and the highest bid *below* x_i (the highest rejected bid) is x_j. If all bidders who bid x_i receive one unit each, then the common price paid by *all* successful bidders (those who bid x_i and above) is $x_j+A_j\delta$ where $A_j>0$ is given explicitly in Theorem 3. If bidders who bid x_i each receive less than one unit (i.e., they share), then the common price is x_i. If the lowest accepted bid is x_T (i.e., $x_i=x_T$), then all successful bidders pay x_T. This auction is a modification of the competitive auction proposed by Vickrey. In the Vickrey auction, the common price is the highest rejected bid. The optimal auction just presented differs from the Vickrey auction in two ways. First it includes a minimum acceptable bid.[12] Second we exploit the assumed discreteness of possible reservation prices by sometimes setting the price above the highest rejected bid. In the theorem below we refer to this auction as the *Modified Vickrey Auction*.[13]

THEOREM 3:

A. *When $Q<N$, an optimal pricing scheme is the Modified Vickrey Auction provided that*

$$(8)\quad x_i+A_i\delta\leqslant\bar{d}\qquad \textit{for } i=T,\ldots,k-1$$

$$\textit{where}\quad A_i=1-\frac{\hat{a}_i-a_i}{b_i}\qquad \textit{for } i=1,\ldots,k-1$$

and the expressions for a_i, $\hat{a}_i$, and b_i are given in the Appendix.

B. *Any buyer with true reservation price at or above x_T will bid his true reservation price when faced with the above auction. Buyers with reservation prices below x_T will choose not to participate.*

[12] Riley and Samuelson consider auctions and derive an optimal minimum bid. Their expression, when specialized to our model, is the same as ours.

[13] Theorem 3 may be viewed as a generalization of Theorem 2 of our earlier paper to more than two agents.

The interpretation of condition (8) in Theorem 3 is that the Modified Vickrey Auction may require a successful bidder to pay more than his endowment for some realizations of the reservation prices. That is for some values of N, Q, and $\bar{d}$, the Modified Vickrey Auction is not feasible, although is optimal whenever feasible. It can be shown that the Modified Vickrey Auction is always feasible if there is only one unit available ($Q=1$). It can also be shown that this auction is feasible if the possible reservation prices (the x_j's) are sufficiently close together (i.e., for sufficiently large k and small δ).[14]

In cases where the Modified Vickrey Auction is feasible (and therefore optimal), Theorems 2 and 3 imply that one might expect to observe either one of two marketing schemes in an environment in which potential demand exceeds capacity. The theory is not rich enough to predict which of these two schemes would be used. Note, however, that the two schemes are quite different. The priority pricing scheme does not require simultaneous participation. For example, one way to implement this scheme is first to offer the product for sale at p_k^*, then gradually lower the price to p_{k-1}^*, etc. In fact we seem to observe this type of scheme in close out sales.[15] On the other hand, the auction scheme requires that all bids be submitted before the allocation of the product and the price are determined. Which scheme one might expect to observe may be determined by such factors as the relative cost of operation which are not included in our analysis.

Finally, note that the two schemes are also different from the point of view of the buyers. Buyers with high reservation prices will prefer the auction scheme while those with low reservation prices will prefer priority pricing (recall that buyers are assumed to know their own reservation prices before the mechanism is chosen). For a buyer with high reservation price, the auction price will, on average, be below the priority price he would choose in the priority pricing scheme. On the other hand, for a buyer with low reservation price, if he purchases the product, the auction price will on average be below his reservation price, but above his priority price.

At this point one might wonder whether the usual *single price scheme* is also optimal. In this scheme, the seller sets a price before the realization of demand, then sells to each buyer who demands the product at that price, up to the limit of his capacity. The single price approach has been analyzed extensively in the literature.[16] In order for this mechanism to be optimal in the present environment, it must yield the same expected profit as the priority pricing scheme. It turns out, however, that when potential demand exceeds capacity, setting a single price yields strictly lower expected profits than priority pricing. To see this, suppose that the best single price is P. Now consider the priority pricing scheme with lowest priority price equal to P. Since $Q<N$, there are, with positive probability, buyers willing to pay more than P in order to get higher priority in the allocation of the good (namely those with reservation prices above P). The priority pricing scheme allows the seller to exploit this fact by charging these buyers higher prices. The quantity sold under either the single price scheme with price P or priority pricing with lowest price P will be the same for each realization of demand. Therefore, the priority pricing scheme with lowest price P is superior. This shows that when $Q<N$, the single price scheme is strictly suboptimal since it does not exploit the potential for discrimination offered by the scarcity of the product.[17]

[14] Some preliminary results suggest that when the number of potential buyers approaches infinity, either all optimal schemes degenerate to single pricing or no common price auction is optimal.

[15] In such a sequential sale, if the buyers are allowed to observe the quantity left at each point in time, then the mechanism is not the priority pricing scheme. Consequently, the outcome might be suboptimal from the seller's point of view. This suggests that the seller might endeavor to conceal remaining stocks.

[16] See, for example, Baron (1971), Crew and Kleindorfer, Holthausen, Leland, Sherman and Visscher, and Tschirhart and Jen.

[17] Note from part B of each theorem (2 and 3) that, after either optimal scheme has been played out, buyers will learn the true reservation prices of other buyers if they are allowed to observe the priority prices paid or the bids of other buyers. One might conjecture, in this case, that discrimination would be thwarted by the attempt of buyers with low reservation prices to resell

TABLE 1—NUMERICAL EXAMPLE

				Price Received	
n_1	n_2	n_3	Probability	*PP*	*MVA*
3	0	0	1/27	3	3
2	1	0	3/27	3-6/7	3-2/3
2	0	1	3/27	4-11/19	3-2/3
1	2	0	3/27	3-6/7	4
1	0	2	3/27	4-11/19	5
1	1	1	6/27	4-11/19	4-5/9
0	3	0	1/27	3-6/7	4
0	2	1	3/27	4-11/19	4-5/9
0	1	2	3/27	4-11/19	5
0	0	3	1/27	4-11/19	5
			$E\pi$	4-1/3	4-1/3

A numerical example might help to clarify the points made above. For this example, we take capacity to be $Q=1$ unit and marginal production costs to be $c=0$. Suppose there are $N=3$ potential buyers each of whom may have a reservation price of \$3, \$4, or \$5 per unit (i.e., $\alpha=3$, $\beta=5$, $k=3$, $\delta=1$). In this case, from (6), the cut-off rank in either the priority pricing scheme or the Modified Vickrey Auction is $T=1$. Thus even a buyer with a reservation price of $x_1=3$ has some probability of receiving some of the good.

We first compute the values of z_j^*. Although these can be computed using the formulae in the Appendix, here we compute them directly as an aid to intuition. Recall that z_1^* is the expected number of units that a buyer with lowest reservation price (\$3) receives in an allocation by rank. Since $Q=1$, the only way a buyer with a reservation price of 3 can obtain any units is if both of the other buyers also have reservation prices of 3. This event has probability $(1/3)(1/3)=1/9$ and if it occurs, each buyer will obtain 1/3 unit. Therefore $z_1^*=(1/9)(1/3)=1/27$, $z_2^*=7/27$, $z_3^*=19/27$, where z_2^* and z_3^* were calculated similarly. Using these values and the formula given in Theorem 2, we calculate the priority prices as $p_1^*=3$, $p_2^*=3-6/7$, $p_3^*=4-11/19$. Therefore, for example, if a buyer is willing to pay p_3^*, he will get one unit if neither of the other two buyers are willing to pay p_3^*, he will get one-half unit if exactly one of the other buyers is willing to pay p_3^*, and he will get one-third unit if both of the other buyers are willing to pay p_3^*. His expected allocation of the good is just z_3^*. Consequently, if a buyer whose reservation price is $x_3=5$ states that he will pay p_3^*, then his expected utility is $(x_3-p_3^*)z_3^*=8/27$. On the other hand, if this buyer stated that he was only willing to pay p_2^*, his expected utility would be $(x_3-p_2^*)z_2^*=8/27$. This demonstrates that any buyer whose reservation price is $x_3=5$ is just willing to pay p_3^* as opposed to p_2^*. The other self-selection properties follow similarly.

With regard to the Modified Vickrey Auction, using Theorem 3, we compute $A_1=2/3$, $A_2=5/9$. This means that if exactly one person bids 5 and at least one other person bids 4, the price will be 4-5/9. If one person bids 5 or 4 and the other two bid 3, the price is 3-2/3. The other cases are shown in Table 1. This table compares the prices received by the seller under priority pricing and the auction schemes as a function of the pattern of true reservation prices of the buyers. The first column gives the assumed pattern of buyers' reservation prices (n_1, n_2, n_3) where n_j is the number of buyers with reservation price x_j. The second column gives the probability which the seller attaches to the pattern given in the first column. The third column gives the price received by the seller in the

their units to buyers with higher reservation prices. This cannot happen, however, since in either scheme, the equilibrium allocation is by rank with rank determined by true reservation price. Thus no buyer receives any units unless all buyers with higher reservation prices are fully satisfied.

priority pricing scheme (*PP*) while the fourth column gives the price received in the Modified Vickrey Auction (*MVA*). At the bottom of columns three and four is shown the expected profit for each scheme.

Note that expected profit in both schemes is 4−1/3. In contrast, if the monopolist were to charge a single price of $4, his expected profits would be 4×(26/27) which is smaller than 4−1/3. Any other single price would yield even smaller expected profits.

B. *Capacity Exceeds Potential Demand* $(Q \geqslant N)$

At the end of the previous subsection, we argued that setting a single price is strictly suboptimal. It is clear from the discussion that the driving assumption was that $Q<N$. It might be conjectured, therefore, that when $Q \geqslant N$, a single price is optimal. In this subsection, we prove this conjecture.

THEOREM 4:

A. *When $Q \geqslant N$, an optimal marketing scheme is to set a single price equal to x_T, where T is defined by* (6). *Each buyer willing to pay x_T (namely those with reservation prices at or above x_T) receives one unit of the product and pays x_T.*

B. *The monopoly price, x_T, is the smallest reservation price such that marginal expected revenue is greater than or equal to marginal cost.*

To gain some insight into this result, suppose $Q \geqslant N$ and the seller announced priority prices and offered higher priority to higher paying customers. If the seller stands ready to sell N units if all N buyers are willing to pay at least the lowest priority price, then higher priority access is meaningless. In this case, clearly all buyers will choose to pay the lowest priority price since they gain nothing from paying a higher price (in either case, they get one unit for sure). In other words, in the absence of scarcity relative to demand, the seller cannot discriminate across buyers based on the probability of receiving the product. Therefore, in order to discriminate, the seller must *create* scarcity by stating that, under no circumstances, will he sell more than say $\hat{Q}<N$ units. In this case, the amount sold will be lower and the price received higher, on average, than in the single price set up. As the theorem shows, the tradeoff is *not* in favor of artificially restricting maximum output.[18]

Note also that the monopoly price x_T will, in general, be such that *less* than N units are sold with positive probability. In particular this will occur whenever some buyers have reservation prices strictly below x_T. Obviously if $x_T=\alpha$ the probability of this event is zero. In general, however, $x_T>\alpha$. The relationship will depend on the cost and demand parameters, as can be seen from (6). Having a monopoly price above α can also be viewed as restricting output. This type of restriction is different from the one discussed in the previous paragraph. There, in order to make discrimination feasible, the seller must restrict output below N independently of the realization of demand. Under a single price scheme the seller stands ready to sell all N units if N buyers have reservation prices at or above x_T. This latter type of output restriction simply reflects the usual "marginal revenue=marginal cost" condition, as is shown in part B of Theorem 4.

Finally the theorem can be interpreted as showing that in monopoly environments with stochastic demand and no binding capacity restriction, an *ex ante* price-setting strategy is superior to an *ex ante* quantity-setting strategy.[19] Most previous studies analyzed the problem of monopoly behavior under uncertain demand by assuming that the monopolist chooses a price *ex ante*. See, for example, Baron (1971), Crew and Kleindorfer, Leland, Meyer, and Sherman and Visscher. Our analysis thus justifies this approach provided capacity exceeds potential demand. Previous studies have, however, employed the price-setting assumption also in

[18] Our model assumes that Q is known by everyone. If Q were, say, random with some probability of being below N, this argument would not go through, and some form of priority pricing would be optimal. We are grateful to George Borts for pointing this out to us.

[19] In fact, Theorem 4 implies only that price setting is weakly superior to quantity setting. The strict superiority (in a limiting case) follows from the result of Section V.

cases where rationing may be required. Our results indicate that in such situations one would expect a monopolist not to use a simple *ex ante* price-setting strategy, but instead use either priority pricing or an auction.

V. Optimal Pricing Schemes with Endogenous Capacity

In this section we suppose that the monopolist can choose his capacity. Capacity must be chosen before demand is known. From the results of Section IV, we know that if capacity is chosen to be less then potential demand, then a priority pricing scheme is optimal whereas if capacity is chosen to exceed potential demand, then a single price scheme is optimal. Thus the problem of choosing capacity can be reduced to the question of whether to choose $Q=N$ (obviously $Q>N$ is pointless) or $Q<N$ and if $Q<N$ is optimal what is the optimal Q.

Clearly, if increasing capacity is free, then the case of endogenous capacity choice is equivalent to having an exogenous capacity limit which is never binding. This is precisely the case analyzed in Section IVB. It therefore follows immediately that if increasing capacity is free, then choosing $Q=N$ and using the single price scheme is optimal. If, however, capacity is at all costly,, a *necessary condition* for a single price with $Q=N$ to remain optimal, is that expected profits are *strictly* higher under monopoly pricing with $Q=N$ than under priority pricing with $Q<N$ *for any* $Q<N$. In the next theorem we show that this condition is in fact satisfied for the limiting case in which the distribution of reservation prices approaches the continuous uniform distribution on the interval $[\alpha, \beta]$. The proof proceeds by taking limits of expected profits as k, the number of equally spaced reservation prices in the interval $[\alpha, \beta]$, approaches infinity.

THEOREM 5: *If capacity can be costlessly chosen, then for the limiting case in which k approaches infinity, the single price* x_T *with* $Q=N$ *is strictly superior to any mechanism with* $Q<N$, *for any* $Q<N$.

Since we show *strict* superiority of a single price scheme with $Q=N$ when capacity is *costlessly* chosen, it follows that even if there is a cost of choosing capacity, a single price is still optimal provided that this cost is sufficiently small. Clearly, if capacity is sufficiently costly, priority pricing will be superior with some $Q<N$.

VI. Conclusions

In this paper, we have endogenously derived optimal monopoly pricing schemes. We found that a crucial aspect of the environment for determining an optimal pricing scheme is capacity limitations. In particular, the policy of charging a single price on a take-it-or-leave-it basis is optimal if and only if capacity restrictions are unimportant. On the other hand, when capacity restrictions are important a more complicated scheme which we call "priority pricing" is optimal. We also find that for some environments with potentially binding capacity, a certain common price auction is also optimal. It is argued in the paper that these results are consistent with casual observations.

Appendix: Proofs of Theorems 2, 3, 4, 5

For the sake of brevity, these proofs are only sketched here. More detailed proofs are available from us on request.

Derivation of $z_T^*, \dots, z_k^*$ *for* $Q<N$:

The appropriate choice of $z_T^*, \dots, z_k^*$ was discussed in the text following equation (6). It can be shown that this procedure for choosing z_j^* results in the following formulae:

$$\text{(A1)} \quad z_i^* = \sum_{m=1}^{i-1} (a_m + b_m) + \hat{a}_i \text{ for } i=T,\dots,k$$

where

$$\text{(A2)} \quad a_i = \varphi^{N-1} \sum_{j=Q}^{N-1} \sum_{l=0}^{Q-2} \binom{N-1}{j} \binom{j}{l} (k-i)^l (i-1)^{N-j-1} \qquad \text{for } i=1,\dots,k-1$$

$$\text{(A3)}\quad \hat{a}_i=\varphi^{N-1}\sum_{j=Q}^{N-1}\sum_{l=0}^{Q-1}\binom{N-1}{j}\binom{j}{l}$$

$$(k-i)^l(i-1)^{N-j-1}\frac{Q-l}{j-l+1}\quad \text{for } i=1,\ldots,k$$

$$\text{(A4)}\quad b_i=\varphi^{N-1}\binom{N-1}{N-Q}(k-i)^{Q-1}$$

$$\left[i^{N-Q}-(i-1)^{N-Q}\right]\quad \text{for } i=1,\ldots,k-1$$

Evaluating the right-hand side of (5) using (A1) and recalling that $z_1^*=z_2^*=\ldots=z_{T-1}^*=0$, we obtain the following expression for the case $Q<N$:

$$\text{(A5)}\quad E\pi=N\varphi\sum_{i=T}^{k}\sum_{m=1}^{i-1}Z_i[(a_m+b_m)+\hat{a}_i]$$

where $Z_i=x_1+(2i-k-1)\delta-c$. Note that (A5) gives the value of any solution of $[P']$.

Proof of Theorem 2:

Consider the allocation

$$\text{(A6)}\quad d_1^*(x_i,R_{-1})=\bar{d}-p_i^*q_1^*(x_i,R_{-1})$$

$$\text{(A7)}\quad q_1^*(x_i,R_{-1})=\begin{cases}0 \text{ if } \sum\limits_{m=i+1}^{k} n_m(R)\geqslant Q \text{ or } i<T\\ \min\left(1,\dfrac{Q-\sum\limits_{m=i+1}^{k}n_m(R)}{n_i(R)}\right) \text{ otherwise}\end{cases}$$

where $n_m(R)$=number of elements of R which equal x_m for $m=1,\ldots,k$, and where p_i^* for $i\geqslant T$ is given in the statement of the theorem and, for notational convenience, $p_i^*=0$ for $i<T$. Let d_j^* and q_j^* for $j>1$ be determined from the symmetry conditions given in the statement of Lemma 1. It can be shown that

$$\text{(A8)}\quad E_1q_1^*(x_i,R_{-1})=\sum_{m=1}^{i-1}(a_m+b_m)+\hat{a}_i=z_i^*$$

It is not difficult to verify using (A8) that this allocation is feasible for $[P]$ and that the objective function evaluated at this allocation equals the right-hand side of (A5). This shows that $[d^*,q^*]$ solves $[P]$. Part B of the theorem is clear since $[d^*,q^*]$ satisfies (SS).

Proof of Theorem 3:

The proposed solution for q_1^* is given in (A7) and now

$$d_1^*(x_i,R_{-1})=\bar{d}-w(R)q_1^*(x_i,R_{-1})$$

where

$$w(R)=\begin{cases}R_{(Q)}\text{ if } R_{(Q+1)}=R_{(Q)}\geqslant x_T\\ R_{(Q+1)}+\delta A_{m[R_{(Q+1)}]}\\ \text{if } R_{(Q)}>R_{(Q+1)}\geqslant x_T\\ x_T\text{ if } R_{(Q+1)}<x_T\end{cases}$$

$R_{(t)}=t$th highest value among the R_i, $i=1,\ldots,N$, for $t=Q,Q+1$ and for any $x_n\in X$, $m(x_n)=n$. The expressions for the A_n are given in the statement of the theorem. It can be shown that $A_n>0$ for $n=1,\ldots,k-1$.

The expected utility of agent 1 for this allocation when $R_1=x_i$ is, again using (A8) and also the definitions of a_n, $\hat{a}_n$ and b_n,

$$\text{(A9)}\quad \bar{d}-E_1w(x_i,R_{-1})q_1^*(x_i,R_{-1})+x_iz_i^*$$

$$=\begin{cases}\bar{d}+x_iz_i^*-\sum\limits_{n=T}^{i-1}a_nx_n\\ \quad-\hat{a}_ix_i-\sum\limits_{n=T}^{i-1}b_n(x_n+A_n\delta)\\ \quad-x_T\sum\limits_{n=1}^{T-1}(a_n+b_n)\text{ if } i\geqslant T\\ \bar{d}\text{ if } i<T\end{cases}$$

The remainder of the proof follows the same strategy as that of Theorem 3 using (A8) and (A9).

Proof of Theorem 4:

When $Q\geqslant N$, it is clear that maximizing the expression on the right-hand side of (5) involves choosing $q_1(x_i, R_{-1})=1$ for $i \geqslant T$, i.e., choosing $z_T^* = \ldots = z_k^* = 1$. Therefore the objective function is

$$\text{(A10)}\quad E\pi = N\varphi \sum_{i=T}^{k} [x_1 + \delta(2i-k-1) - c]$$

$$= N\varphi(k-T+1)(x_T - c)$$

The proposed solution for this theorem is

$$d_1^*(x_i, R_{-1}) = \bar{d} - x_T q_1^*(x_i, R_{-1})$$

$$q_1^*(x_i, R_{-1}) = \begin{cases} 1 & \text{for } i \geqslant T \\ 0 & \text{for } i < T \end{cases}$$

The remainder of the proof of part A follows the same pattern as in Theorems 2 and 3.

To prove part B, the change in expected revenue ΔER_i when the monopoly price changes from x_i to x_{i+1} is given by $\Delta ER_i = (N/k)[(k-i+1)x_i - (k-i)x_{i+1}] = (N/k)(2x_i - x_k)$. The change in expected quantity sold is $(N/k)[(k-i+1)-(k-i)] = N/k$. Therefore the marginal expected revenue is $2x_i - x_k$ which is greater than or equal to marginal cost c if and only if

$$\text{(A11)}\qquad x_i \geqslant \tfrac{1}{2}(x_k + c)$$

But as mentioned above, x_T is the smallest x_i such that (A11) holds. This proves part B of the theorem.

Proof of Theorem 5:

Clearly it suffices to show that in the limit as $k \rightarrow \infty$, expected profits using a single price x_T with $Q=N$ is strictly greater than (A5) for any $Q<N$.

From (6), it can be shown that

(A12)

$$t_\infty \equiv \lim_{k\rightarrow\infty} (T/k) = \max\left[0, \frac{c-2\alpha+\beta}{2(\beta-\alpha)}\right] < \infty$$

Now, denote by $\pi_M(k)$ the value of expected profits under a single price x_T when $Q=N$. From the proof of Theorem 4

(A13)

$$\lim_{k\rightarrow\infty} \pi_M(k) = \lim_{k\rightarrow\infty} \frac{N}{k}(k-t+1)(x_T - c)$$

$$= \begin{cases} N(\alpha - c) & \text{if } c < 2\alpha - \beta \\ \dfrac{N(\beta - c)^2}{4(\beta - \alpha)} & \text{if } c \geqslant 2\alpha - \beta \end{cases}$$

Next, denote by $\pi(Q,k)$ the value of expected profits for any $Q<N$ given in (A5). Thus

$$\pi(Q,k) = N\varphi\left\{(x_1 - c)\sum_{i=T}^{k} z_i^* \right.$$

$$\left. -(k+1)\delta \sum_{i=T}^{k} z_i^* + 2\delta \sum_{i=T}^{k} i z_i^* \right\}$$

It can be shown, using the above expression that $\lim_{k\rightarrow\infty}\pi(Q,k) < \lim_{k\rightarrow\infty}\pi_M(k)$.

REFERENCES

D. Baron, "Demand Uncertainty in Imperfect Competition," *Int. Econ. Rev.*, June 1971, *12*, 196–208.

______, "Incentive Contracts and Competitive Bidding," *Amer. Econ. Rev.*, June 1972, *62*, 384–94.

G. Brown, Jr. and M. B. Johnson, "Public Utility Output and Pricing Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

M. A. Crew and P. R. Kleindorfer, "Reliability and Public Utility Pricing," *Amer. Econ. Rev.*, Mar. 1978, *68*, 31–40.

M. Harris and A. Raviv, "Allocation Mechanisms and the Design of Auctions," *Econometrica*, forthcoming.

______ **and R. M. Townsend,** "Resource Allocation under Asymmetric Information," *Econometrica*, Jan. 1981, *49*, 33–64.

C. Holt, "Competitive Bidding for Contracts Under Alternative Auction Procedures," *J. Polit. Econ.*, June 1980, *88*, 433–45.

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

H. E. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

M. G. Marchand, "Pricing Power Supplied on an Interruptible Basis," *European Econ. Rev.*, Oct. 1974, *5*, 263–74.

E. S. Maskin and J. E. Riley, "Price Discrimination and Bundling: Monopoly Selling Strategies when Information is Incomplete," mimeo., Univ. California-Los Angeles 1979.

S. Matthews, "Information Acquisition in Discriminatory Auctions," working paper, Univ. Illinois, Mar. 1979.

R. A. Meyer, "Monopoly Pricing and Capacity Choice Under Uncertainty," *Amer. Econ. Rev.*, June 1975, *65*, 326–37.

R. Myerson, "Optimal Auction Design," *Math. Operations Res.*, Feb. 1981, *6*, 58–73.

______, "Incentive Compatability and the Bargaining Problem," *Econometrica*, Jan. 1979, *47*, 61–74.

J. Riley and W. Samuelson, "Optimal Auctions," *Amer. Econ. Rev.*, June 1981, *71*, 381–92.

R. Sherman and M. Visscher, "Second Best Pricing with Stochastic Demand," *Amer. Econ. Rev.*, Mar. 1978, *68*, 41–53.

J. E. Stiglitz, "Monopoly, Non-Linear Pricing and Imperfect Information: The Insurance Market," *Rev. Econ. Studies*, Oct. 1977, *45*, 407–30.

J. Tschirhart and F. Jen, "Behavior of a Monopoly Offering Interruptible Service," *Bell J. Econ.*, Spring 1979, *10*, 244–58.

W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J. Finance*, Mar. 1961, *16*, 8–37.

R. B. Wilson, "Auctions of Shares," working paper, Grad. School Business, Stanford Univ., 1979.

A Pure Theory of Strategic Behavior and Social Institutions

By Earl A. Thompson and Roger L. Faith*

The dominant economic models of the interaction between maximizing individuals, the models of Cournot and Stackelberg, assume what economists have come to call "nonstrategic behavior." The individuals represented in such models cannot make and communicate prior commitments to reaction functions in order to influence the subsequent decisions of others. Yet the importance of strategic behavior in achieving realistic solutions to problems of individual interaction is becoming increasingly apparent. Thomas Schelling, in his 1963 classic, and several subsequent writers on bilateral bargaining gave numerous examples in which prior reaction commitments are required to produce realistic solutions to bilateral bargaining problems. James Buchanan constructed examples in which prior reaction commitments are required to prevent charitable donations from worsening the conditions that the donors wish to improve. We have shown in our 1979 paper that observed firm interaction in modern, concentrated industries can be understood only by assuming that some producers have precommitted themselves to certain reactions to the outputs of others.

More basic examples concern the institutions of exchange and property rights. Rational exchange cannot exist under Cournot or Stackelberg interaction in a finite horizon trading model. Under Stackelberg interaction, the last deliverer in an exchange, or series of exchanges, having already obtained all he ever will from the others, has no incentive whatever to deliver; so any prior deliveries by others would also be irrational. Under Cournot interaction, where no party observes the actual deliveries of others (there are only estimates of the deliveries of others, estimates which are correct in equilibrium), no party has an incentive to deliver as one's receipts are independent of his deliveries. Rational exchange in a finite horizon model —and a typical exchange in the real world— requires a prior commitment that imposes greater costs for nondelivery than the goods are worth. Similarly, the existence of private property itself generally requires prior commitments to retaliate against potential aggressors, such retaliation generally requiring expenditures by protectors in excess of the value of the property (see Thompson, 1979).

The primary purpose of this paper is to develop an *n*-person model of strategic behavior for the "pure" case in which no individual suffers any direct costs of committing himself to or communicating any one of his possible reaction functions. Both the basic model, which also assumes a finite set of social alternatives, and the model's solution, are specified in Section I. Our central result, the Pareto optimality of the solution under strict preference relations, is demonstrated in Section II. (An exception arising under weak preference relations is discussed at length in our 1980 paper.)

Section III is mainly pedogogical; it contrasts our solution to the standard noncooperative solution to a prisoner's dilemma game in order to clarify possible ambiguities concerning the nature of our communication and commitment assumptions.

It is tempting to apply our central optimality result to the interactions between subgroups of communicating individuals in environments containing outside enforcers. In particular, it is tempting to infer that we are establishing the traditional conjecture that any subset of individuals, if they perfectly communicate with one another, will interact so as to achieve a joint optimum among

*University of California-Los Angeles and Virginia Polytechnic Institute and State University, respectively. We benefitted substantially from comments on earlier drafts by Ron Batchelder, Jack Hirshleifer, Louis Makowski, Joe Ostroy, Lloyd Shapley, and a referee of this journal.

themselves. But there is no reason to suppose that the outside enforcers would not affect the reaction functions of the insiders, thereby violating the condition that the individuals freely select from all possible reaction functions. Furthermore, even if the outside enforcers induced no alterations in insider decisions, the fact that certain insiders, in general, devote overhead resources to establishing the priority of their individual commitments makes it possible for outsider intervention to benefit everyone by inducing reductions in such resource expenditures. If, however, these resource costs were somehow prevented, and if the only other imposition on the subgroup is a compensatory, common-law property rights system, then the so-constrained solution would be a Pareto optimum for the subgroup (see our 1980 article). This amounts to a formalization of the Coase Theorem. The present paper, however, keeps the analysis at the level of the entire social group, wherein no outside imposition on feasible reaction functions can appear to threaten the optimality result and, because of the absence of outsiders and therefore an absence of anyone to whom we can appeal to gain an allocative correction, the resource losses in establishing a prior commitment become *unavoidable* deadweight costs. Correspondingly, our optimality result becomes a positive hypothesis rather than a normative statement, the conclusion being that, except for unavoidable overhead costs, equilibrium allocations, which always exist in our model, will be Pareto optimal. Alternatively, since the "institutions" that an individual faces are defined by the reactions of others to his own actions, our positivistic central conclusion states that equilibrium institutions always exist and are Pareto optimal.

The extreme assumptions regarding the commitment-making and communication abilities of our individuals mask the empirical relevance of the model. Section IV introduces plausible restrictions on the physical environment that allow us to reduce the information and commitment-making assumptions of the general model to plausible levels. The resulting special model amounts to a theory of social institutions with sufficient empirical power to suggest explanations for broad variations, historical and cross-sectional, in observed political and economic institutions.

John von Neumann and Oscar Morgenstern, in their pathbreaking work on the theory of games, also argued that the ultimate aim of their exercise was to determine institutions endogenously. Indeed, our basic model in Section I is merely a von Neumann-Morgenstern "perfect information game" played over strategies rather than simple actions (or "plays of a game"). While von Neumann and Morgenstern explicitly recognized (Sec. 11.3) that games could be constructed in which strategies are communicated in the same way as the actions in their perfect information games, they saw nothing novel about such games. For such games posed no new problem in the development of solution concepts or the existence of solutions. Perhaps, had they been more interested in evaluating the Pareto optimality of solution outcomes or in abstractly characterizing basic social institutions such as private property or contracts, they would have devoted some of their prodigious intellectual resources to characterizing games with perfect information concerning strategies as well as actions. But perhaps not too. For such games represent an uncomfortable hybrid within the corpus of game theory in that the games are, strictly speaking, neither "cooperative" nor "noncooperative." The games allow more information than noncooperative games in that they explicitly allow preplay communication between the players. Yet the games are not cooperative either in that they contain no exogenous "characteristic functions" mysteriously assigning payoffs to "coalitions" of players and, correspondingly, no prior imposition of group rationality conditions. Our model is therefore distinct from other game-theoretic models in that it contains a theory of individually rational communication and, correspondingly, a theory of individually rational cooperation.[1]

[1] This is not to say that conventional cooperative game theory cannot be reformulated to produce a game-theoretic model similar to our own generalization of Schelling's bargaining model. Indeed, such a reformulation has been recently achieved by Rosenthal.

Section V identifies the source of the difference between our theory and conventional cooperative game theory. In so doing, it exposes a basic defect in the latter as a theory of strategic communication. The general argument is then applied to voting processes. Perfect strategic communication is shown to preclude the "voters' paradox" that underlies modern, Arrowian critiques of democracy, critiques implicitly dependent upon viewing the choice process as a conventional cooperative game (see Robert Wilson).

I. The Basic Model and Its Solutions

A. *The Physical Environment*

An individual is denoted $i, i=1,\ldots,n$. An action of individual i is denoted x_i, where $x_i \in X_i$, a finite set of feasible actions of individual i. A possible social choice, or *allocation*, is defined by an n-dimensional set of actions, and is denoted $x=(x_1,x_2,\ldots,x_n)$, so that $x \in \Pi_{i=1}^{n} X_i$. To describe individual preferences, each individual i is given a complete, transitive, irrelexive, antisymmetric, binary relation $\succ_i$, defined over $\Pi_{i=1}^{n} X_i$. This description, in effect, assumes away indifference between any pair among the finite set of possible allocations. The motivation for this assumption and the effects of indifference on our central results will be discussed later. A Pareto optimum is an allocation, x', $x' \in \Pi_{i=1}^{n} X_i$, for which there is no alternative allocation, x'', $x'' \in \Pi_{i=1}^{n} X_i$, such that $x'' \succ_i x'$ for all i. Several Pareto optima may exist.

B. *Institutional Possibilities*

The institutions facing an individual can be completely described by the reactions of other individuals to his own actions. But institutions or reactions are not taken here as given; they are derived. This is done by allowing each individual to select, among all feasible reaction functions, a function which is maximal with respect to his preference relation. But we want individuals to *know* the institutions and thus the reaction functions of others. And for this to hold generally, the functions must be communicated in sequence. Thus, for the individuals to know the institutions, the first communicator, say individual 1, presents the reaction function,

$$(1)\qquad x_1 = f_1(x_2,\ldots,x_n)$$

to the other individuals; the second communicator, say individual 2, then presents

$$(2)\qquad x_2 = f_2(x_3,\ldots,x_n)$$

to individuals 3 through n; the third communicator then presents

$$(3)\qquad x_3 = f_3(x_4,\ldots,x_n)$$

to individuals 4 through n, and so on up to the $n-1$st communicator, who presents

$$(4)\qquad x_{n-1} = f_{n-1}(x_n)$$

to the nth individual, who has no need to communicate. Once the action of the nth individual is taken, the action of the $n-1$st individual is determined. Once this pair of actions is taken, the action of individual $n-3$ is determined, and so on up until an allocation is determined as a chain reaction from the nth individual's action. The set $(f_1, f_2,\ldots,f_{n-1})$ is thus a complete institutional description. The feasible choice set, or strategy set, of individual 1 is the set of *all* functions from $\Pi_{i=2}^{n} X_i$ to X_1. This can be represented by the functional variable F_1. Similarly, $F_2,\ldots,F_{n-1}$ can be used to represent the respective strategy sets of individuals 2 to $n-1$. The product space $\Pi_{i=1}^{n-1} F_i$ thus represent the world's institutional possibilities. The strategy set of individual n is X_n.

A question may arise as to why some individuals do not present reaction functions to other individuals who are higher up in the communication hierarchy. Consider individual n. Facing the prior strategies of the other $n-1$ individuals, he sees that the eventual allocation must be consistent with the chosen reaction functions of each of the $n-1$ prior selectors. Hence, if individual n responds to the prior selectors with a simple action, he will have a free choice over all allocations consistent with the prior reaction functions.

But if n responds with a function of prior actions, thus giving further choices to the prior strategy selectors, he can only reduce his original choice out of the same set of possible allocations. He cannot expand the set of possible outcomes because any eventual outcome must be consistent with the given $n-1$ reaction functions. Similarly, if the $n-1$st strategy selector presents a reaction function rather than an action to his prior strategy selectors for a given action of individual n, he is giving them the choice of actions consistent with the set of reaction functions he faces and thus can be no better off. This also applies, in like fashion, to individuals $n-2$ to 2, so that it is in no individual's interest to present a reaction function to a prior strategy selector.

The above world, which can now be viewed as a "game," differs from the standard, von Neumann-Morgenstern, "perfect information" (and majorant-minorant) games in that some individuals are allowed to communicate their strategies to others before the latter select their own strategies. Thus, in the von Neumann-Morgenstern world, a player will not adopt a special strategy in order to influence the subsequent strategies (and actions) of others simply because he cannot communicate it and therefore cannot use it to influence the subsequently chosen strategies. In contrast, in the above world, each of the first $n-1$ players communicates his strategy to all subsequent strategy selectors. And response strategies of the subsequent selectors are known a priori by the prior strategy selector because they are the rational responses to the given strategy of the prior selector.

Due to the additional information implied by having individuals communicate committed reaction functions to subsequent players before the latter select their strategies and before any actions are taken, the information implied by the above sequence of reaction functions is called "truly perfect." More specifically, under "truly perfect information," each player: 1) knows with certainty the reaction functions chosen prior to his strategy choice; 2) can freely choose (i.e., commit himself to) and communicate any reaction function consistent with $x \in \Pi_{i=1}^{n} X_i$ and prior reaction functions; and 3) knows with certainty the rational responses to each of his possible reaction functions by all subsequent strategy selectors.

An alternative formulation of the above model, one which obviates the communication and commitment concepts used above, is to allow players in an n-person perfect information game with n moves an earlier, *additional* series of $n-1$ special moves in which each can perform a new kind of action, one which has the effect of reducing his subsequent feasible responses to the regular actions of others to a single, specified regular action. This formulation, a case of which is outlined in Schelling, produces the same result as above but adds a cumbersome, logically unnecessary step to the formal development.

While Nigel Howard has produced a general class of games (called "*jk*-metagames") containing strategies contingent on the strategies of other strategy selectors, he does not assume truly perfect information. Correspondingly, he does not adopt a perfect information solution concept. Rather, he adopts, without substantive justification, the von Neumann-Morgenstern and Nash "no-regret" solution concept in which each strategy selector accepts *as given* the strategies of all other strategy selectors. This amounts, as Howard recognizes, to assuming uniformly zero information regarding the strategies of others at the time of strategy selection. For if the choice of a strategy selector were perceived by subsequent strategy selectors, it would, in general, influence the latter's selections. Such games, besides being theoretically unsatisfying in that they typically generate a multiplicity of solution points, some of which are optimal and others nonoptimal (see Howard, p. 58), are empirically unsatisfying in that observed commitments are, as pointed out in the introduction, typically communicated to others in order to influence their strategy selections.

C. *Equilibrium Institutions, or "Solutions," under Truly Perfect Information*

A solution, $(f_1^*, \ldots, f_{n-1}^*, x_n^*)$, is a set in which the ith variable is maximal with respect to $\succ_i$ for given values of $f_1,\dots, f_{i-1}$. A solution can be constructed as follows: First, we find, for individual n, x_n^*, the point in X_n such that, for all $x_n \neq x_n^*$,

$$\{f_1(f_2,\dots, f_{n-1}, x_n^*),$$
$$f_2(f_3,\dots, f_{n-1}, x_n^*),\dots, x_n^*\} \succ_n$$
$$\{f_1(f_2,\dots, f_{n-1}, x_n),$$
$$f_2(f_3,\dots, f_{n-1}, x_n),\dots, x_n\}$$

This solution determines a dependency of x_n^* on $f_1, f_2, \dots, f_{n-1}$, which we write $x_n^*[f_1,\dots, f_{n-1}]$. Then, for individual $n-1$, we find a reaction function, f_{n-1}^*, such that for all $f_{n-1} \in F_{n-1}$, $f_{n-1} \neq f_{n-1}^*$,

$$\{f_1(f_2,\dots, f_{n-2}, f_{n-1}^*, x_n^*[f_1,\dots, f_{n-2}, f_{n-1}^*])$$
$$,\dots, f_{n-2}; f_{n-1}^*, x_n^*[f_1,\dots, f_{n-2}, f_{n-1}^*]\}$$
$$\succ_{n-1}\{f_1(f_2,\dots, f_{n-2}, f_{n-1},$$
$$x_n^*[f_1,\dots, f_{n-2}, f_{n-1}]),\dots, f_{n-2}, f_{n-1},$$
$$x_n^*[f_1,\dots, f_{n-2}, f_{n-1}]\}$$

This solution determines the dependency of f_{n-1}^* on $f_1, f_2,\dots,$ and f_{n-2}, which we describe as $f_{n-1}^*[f_1,\dots, f_{n-2}]$. Then, for individual $n-2$, we find a reaction function, f_{n-2}^*, such that, for all $f_{n-2} \in F_{n-2}$, $f_{n-2} \neq f_{n-2}^*$,

$$\{f_1(f_2,\dots, f_{n-2}^*, f_{n-1}^*[f_1,\dots, f_{n-2}^*],$$
$$x_n^*[f_1,\dots, f_{n-2}^*, f_{n-1}^*[f_1,\dots, f_{n-2}^*]]),$$
$$\dots, f_{n-2}^*, f_{n-1}^*[f_1,\dots, f_{n-2}^*],$$
$$x_n^*[f_1,\dots, f_{n-2}^*, f_{n-1}^*[f_1,\dots, f_{n-2}^*]]\}_{n-2}$$
$$\{f_1(f_2,\dots, f_{n-2}, f_{n-1}^*[f_1,\dots, f_{n-2}],$$
$$x_n^*[f_1,\dots, f_{n-2}, f_{n-1}^*[f_1,\dots, f_{n-2}]]),$$
$$\dots, f_{n-2}, f_{n-1}^*[f_1,\dots, f_{n-2}],$$
$$x_n^*[f_1,\dots, f_{n-2}, f_{n-1}^*[f_1,\dots, f_{n-2}]]\}$$

This solution thus determines the dependency of f_{n-2}^* on $f_1, f_2,\dots,$ and f_{n-3}, which we write as $f_{n-2}^*[f_1,\dots, f_{n-3}]$. The process continues until we have determined f_1^*. Since f_1^* does not depend on any prior functions, we can use it to determine the succeeding reaction functions by successively substituting starred values into $f_2^*[f_1]$, $f_3^*[f_1, f_2],\dots,$ and $f_{n-1}^*[f_1, f_2,\dots, f_{n-2}]$. In this way, a solution $(f_1^*, f_2^*,\dots, x_n^*)$, which implies a solution allocation $(x_1^*, x_2^*,\dots, x_n^*)$, is determined.

The finite structure of the successive maximization problems, along with the completeness and transitivity of $\succ_i$, assures us that a solution *always* exists.

D. *Determination of Priority and the Role of Commitments*

The above game, with its predetermined priority, is not symmetric in that its solution will generally vary with the order of priority. While one may think of the priority in strategy making in the above model as being arbitrarily determined by an umpire of the game or some random device, it is much more realistic to determine the order of strategy selection in a higher-order game.

Such higher-order games come in two forms. One form corresponds to a world containing a higher authority, an outside player who assigns hierarchical positions according to competitive bids for the positions. In such a world, the outside player may also serve the function of an enforcer who assigns hierarchal position and punishes any player who does not carry out his announced strategy. This could eliminate insider resources devoted to establishing prior commitments and also assure the unrestricted commitment ability which characterizes the basic model, thereby preventing the inefficiency possibilities raised in the introduction. The game of contracting with outside parties to guarantee commitments, the higher-order game which determines the order of strategy selection, and the game described in the above sections may all be combined into a single game in which players interact to determine the method of enforcement, the order, and the specific form of all strategy commitments. (See our 1979 paper for a specification of such a game and an existence proof for the game.) Since a player in such a game can lower the bids against him for a given position in the hierarchy by choosing a strategy that yields more benefits to his competing subordinates, the solutions to such a game differ somewhat from those described above (again, see our paper). So our optimality theorem does not generally apply to these games. Whether outside authorities establish Pareto optimal rules cannot be theoretically determined when there is an outside authority.

The second form of higher-order game is a warlike affair with no higher authority. Pareto inefficient, Nash-von Neumann and Morgenstern, noncooperative games apply in determining the order of strategy selection. But the solution characteristics of our own lower-order game are unaffected by the higher-order game. War losses are strictly sunk costs once a hierarchy is established and our game is ready to be played. Hence, we shall concentrate our applications of the model on raw states of the world, where outside authorities do not exist (Section IV).

E. *Lack of Realism*

It is grossly unrealistic to arrange *all* of the individuals in any observed society into a hierarchy of strategy selectors in which each must receive the strategies of the previous selectors before transmitting his own strategy to others. Furthermore, it is similarly unrealistic to give all (but one) individuals the ability to adopt strategies which commit them to carrying out actions which, at the time of their undertaking, may be irrational.

However, as shall be shown in Section IV, under certain, rather realistic, specializing assumptions regarding the physical environment, the model does not require either of these extreme characteristics.

II. Pareto Optimality

Besides unqualified existence, the solution has the important property of Pareto optimality. That is, institutions formed under truly perfect information always imply Pareto optimal allocations.

To prove this, suppose the solution allocation $(x_1^*,\dots,x_{n-1}^*,x_n^*)=x^*$ is not Pareto optimal. Then there is a point, $x^0=(x_1^0,\dots,x_n^0)\in\Pi_{i=1}^n X_i$ such that $x^0\succ_i x^*$ for all i. A set of reaction functions generating x^0 as an allocation is given by $(f_1^0,\dots,f_{n-1}^0)$. Of course, $(f_1^*,\dots,f_{n-1}^*,x_n^0)\neq(f_1^0,\dots,f_{n-1}^0,x_n^0)$; otherwise x^0 would be the solution. Now let individual 1 consider:

$$(5)\quad f_1(f_2,\dots,f_{n-1},x_n)=\begin{cases} f_1^0 \text{ if } (f_2,\dots,f_{n-1},x_n) \\ \quad =(f_2^0,\dots,f_{n-1}^0,x_n^0) \\ f_1^* \text{ otherwise}\end{cases}$$

This may induce each subsequent strategy selector to reorder his strategy in $f_2^0,\dots,f_{n-1}^0,x_n^0$ relative to $f_2^*,\dots,f_{n-1}^*,x_n^*$. However, as it does not alter the allocations resulting from nonsolution strategies *other than* $f_2^0,\dots,f_{n-1}^0,x_n^0$, it does not alter anyone's ordering of these other strategies relative to $f_2^*,\dots,f_{n-1}^*,x_n^*$. Therefore, because $x^0\succ_1 x^*$, individual 1 is no worse off under (5) than under his original strategy.

We next let individual 2 consider, in view of (5),

$$(6)\quad f_2(f_3,\dots,f_{n-1},x_n)=\begin{cases} f_2^0 \text{ if } (f_3,\dots,x_n) \\ \quad =(f_3^0,\dots,x_n^0) \\ f_2^* \text{ otherwise}\end{cases}$$

This similarly cannot hurt individual 2. We continue on to individual n, who now faces (5), (6),... . Thus, $(f_1^0,\dots,f_{n-1}^0,x_n^0)\Rightarrow x^0$will result if he picks $x_n=x_n^0$; and $(f_1^*,\dots,f_{n-1}^*,x_n^*)$ if he picks his solution action. Since $x^0\succ_n x^*$, he picks the former. The supposition that there is a Pareto nonoptimal solution is thus immediately contradicted: For the supposition implies that the players individually prefer a nonsolution set of strategies $(f_1^0,\dots,f_{n-1}^0,x_n^0)$ to the solution set $(f_1^*,\dots,f_{n-1}^*,x_n^*)$.

III. Contrast to Prisoner's Dilemma Game

The above game contrasts sharply with the familiar prisoner's dilemma game. A prisoner's dilemma payoff matrix is illustrated in Figure 1. In a prisoner's dilemma game with conventional perfect information, the standard von Neumann-Morgenstern (vNM) perfect-information solution applies. Following vNM, the player who has the "second move," say the row player R, has his strategy set expanded beyond the set of simple actions to include that player's possible reactions to the various actions of his opponent. The column player C then has the "first move." The normal form of the game is shown in Figure 2, where, for example, $x'_R|x''_C$ means that R adopts his first action if C adopts his second. Solving the game, C peruses each column to determine which action R will select (i.e., which action maximizes R's payoff) for each of the given actions of C and then selects the action which maximizes his own payoff given the resulting action of R. Player C's optimal strategy, in light of R's rational response, is to play x''_C. This leads R to play x''_R (or $(x'_R|x'_C, x''_R|x''_C)$), resulting in the jointly inefficient outcome (x''_R, x''_C). (Since x''_R is a "dominant strategy" for R, (x''_R, x''_C) also represents a Nash-vNM "no-regret" solution to the normal form, the pair of strategies such that no player can increase his payoff by changing his strategy, *given* the strategies of the other players.) In contrast, under our assumption of strategic communication, R, while moving second, is able to commit himself to a strategy and communicate it to C before C moves. Player R rationally commits himself to $(x'_R|x'_C, x''_R|x''_C)$ in view of C's rational responses to R's various possible strategies. R's strategy thus becomes a committed reaction function rather than a narrowly rational response function. C then rationally chooses x'_C so the solution is the jointly efficient solution (x'_R, x'_C). Hence, as long as the second mover is able to communicate his strategy to the first before the first makes his move, the solution is the jointly efficient outcome.

R \ C	DON'T CONFESS x'_C	CONFESS x''_C
DON'T CONFESS x'_R	4, 4	0, 10
CONFESS x''_R	10, 0	1, 1

FIGURE 1. PAYOFF MATRIX

R \ C	x'_C	x''_C
x'_R	4, 4	0, 10
x''_R	10, 0	1, 1
$x'_R\|x'_C,\ x''_R\|x''_C$	4, 4	1, 1
$x''_R\|x'_C,\ x'_R\|x''_C$	10, 0	0, 10

FIGURE 2. NORMAL FORM OF PRISONER'S DILEMMA

IV. Specializations and Empirical Applications

A. *The Nature of the Solution under an Additional Assumption*

To give some empirical power to the above model, we shall now assume that for any individual below the first strategy selector, there is some individual higher up in the decision hierarchy who "can punish" him. To define this formally, first let x^1 be the allocation which individual 1 prefers to all other x in $\Pi_{i=1}^n X_i$. Individual k is said to "punish" individual j if k selects an $x_k = x_k^{P(j)}$ such that $x^1 \succ_j (x_1, \ldots, x_k^{P(j)}, \ldots, x_n)$ for any $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ in $\Pi_{i \neq k} X_i$. Individual k can punish individual j iff $x_k^{P(j)} \in X_k$. Our assumption, then, is that for *any* j, $j = 2, \ldots, n$, there is a k, $k < j$, such that $x_k^{P(j)} \in X_k$. In other words, for any j, there is someone higher up in the hierarchy that can select an action which makes him worse off than he would be under x^1. The assumption does not appear to be unrealistic once we recognize that, in the real world, almost any healthy adult can inflict damages on almost any *single* other individual to the extent that the victim would prefer serving as a slave to suffering the damages.

The assumption implies that individual 1 can punish individual 2, that individuals 1 or 2 can punish 3, that 1,2, or 3 can punish 4, etc. It follows that individual 1 can, by adopting the appropriate strategy, effect the punishment of *any* other single individual if that individual chooses an action which is not an element of x^1. To see this, let j now represent the individual of lowest hierarchial rank that chooses an action which is not an element of x^1 and let i now be the first individual, going up the hierarchy, who can punish j. If $i = 1$, individual 1 can effect the punishment of j directly. If $i > 1$, let h be the first individual, going up in the hierarchy from i, who can punish i. If $h = 1$, individual 1 can adopt a strategy in which he punishes i if $x_j \neq x_j^1$ and chooses x^1 if $(x_2, \ldots, x_n) = (x_2^1, \ldots, x_n^1)$. If 1 adopts such a strategy, i's optimal strategy is one in which he punishes j if $x_j \neq x_j^1$ and chooses x_i^1 if $(x_{i+1}, \ldots, x_n) = (x_{i+1}^1, \ldots, x_n^1)$. For, given 1's strategy, any strategy of i that does not have this characteristic will, by inducing j to pick an $x_j \neq x_j^1$, generate a punishment action by 1 against individual i and thus an allocation which is worse for him than the alternative, x^1. If $h > 1$, but g, the first person in the hierarchy above h who can punish h, equals 1, then 1 can adopt a strategy in which he punishes h if $x_j \neq x_j^1$ and chooses x^1 if $(x_2, \ldots, x_n) = (x_2^1, \ldots, x_n^1)$. For any such strategy of individual 1, h's optimal strategy is one in which he punishes i if $x_j \neq x_j^1$ and adopts x^1 if $(x_{h+1}, \ldots, x_n) = (x_{h+1}^1, \ldots, x_n^1)$ in view of prior reaction functions which assure him that x^1 will result if $(x_h, \ldots, x_n) = (x_h^1, \ldots, x_n^1)$. If $g > 1$, the same argument applies so that g will be induced by prior strategies to punish h, who will then be induced to punish i if $x_j \neq x_j^1$. Eventually the sequence must reach individual 1 and the punishment accomplished. This applies to any j, where j is again the first player who chooses an action that is not an element of x^1.

Denoting the resulting ordered set of individuals by $(1, a(j), \ldots, g(j), h(j), i(j), j)$, when the set contains only two elements, 1 and j, the second element, $a(j)$, equals j.

Under the punishability assumption described above, individual 1's optimal reaction function is given by

$$x_1 = \begin{cases} x_1^1 \text{ if } (x_2, \ldots, x_n) = (x_2^1, \ldots, x_n^1) \\ x_1^{P(a(j))} \text{ otherwise} \end{cases}$$

Given this, regardless of the reaction functions in between 1's and $a(j)$'s, the optimal reaction functions from $a(j)$ to $i(j)$ become

$$x_{a(j)} = \begin{cases} x_a^1 \text{ if } (x_{a+1}, \ldots, x_n) = (x_{a+1}^1, \ldots, x_n^1) \\ x^{P(b(j))} \text{ otherwise,} \end{cases}$$

$$x_{i(j)} = \begin{cases} x_i^1 \text{ if } (x_{i+1}, \ldots, x_n) = (x_{i+1}^1, \ldots, x_n^1) \\ x_i^{P(j)} \text{ otherwise} \end{cases}$$

If $i(j)$ did not select this strategy so that j failed to adopt x_j^1, then $h(j)$ would, according to the prior functions in the above sequence, punish him. Given this chain of reaction functions, the jth individual, $j < n$, is irrational if he chooses a reaction function other than one in which

$$x_j = x_j^1 \text{ if } (x_{j+1}, \ldots, x_n) = (x_{j+1}^1, \ldots, x_n^1)$$

(Since by definition, players $j+1, \ldots, n$

choose $(x^1_{j+1},\dots, x^1_n)$, j's responses to other actions are irrelevant.) And, of course, if $j=n, j$ is irrational to pick $x_n \neq x^1_n$. This is because the allocation resulting from such deviations is less preferred by individual j. Hence, assuming rationality, as defined in our solution concept: first, n will not diverge from x^1_n in order to save being punished, so $j \neq n$; then $n-1$ will not diverge for the same reason, so $j \neq n-1$; and so on. Since no j exists under the definition of a solution, the solution must be x^1.

The result of our rather plausible, apparently innocent, assumption is thus that any solution to social interaction under truly perfect information is dictatorial in the sense of Kenneth Arrow. The result thus takes the bite out of Arrow's possibility theorem, since dictatorial social welfare functions are always possible under Arrow's reasonability conditions. A social welfare function may be constructed from the preferences of individual 1, who is the dictator, and Arrow's problem vanishes. At the same time, the result provides us with an empirical, rather than metaphysical, foundation for adopting the maximization of a social welfare function as a social goal.

The Arrow theorem will become possibly relevant again once we have our dictator allow subgroups in the population to interact in a nondictatorial fashion. We shall, however, eventually find a general inconsistency between Arrow's implicitly cooperative game (Wilson) and *any* game with truly perfect information.

The punishability condition—when complemented by another rather realistic specialization described below—will also allow us to substantially reduce the informational requirements of our model.

B. *Reducing the Informational Requirements of the Model*

As mentioned earlier (Section I., Part E), it is highly unrealistic to give all individuals (except one) the ability to 1) communicate reaction functions to all subsequent strategy selectors, and 2) select reaction functions that require a prior commitment to narrowly irrational behavior. This lack of realism can now be removed by adding the assumption that the set of individuals $(1, a(j),\dots, j)$ is independent of j for any $j>m$, a relatively small number compared to n. The rest of the individuals, the bulk of the population, seeing that these players enforce x^1, will simply choose their element of x^1 without having to observe the strategies selected by the others in the group. That is, they simply set $x_i = x^1_i$ and do not exhibit any punishment actions, thus producing a set of degenerate, constant reaction functions, $x^1_i, i>m$. This greatly reduces the information requirement of our solution in that the large majority of the population need neither communicate their reaction functions to the others nor commit themselves to narrowly irrational reactions.

As mentioned in Section I, when there is no predetermined outside authority to assign a hierarchy and enforce commitments to the announced reaction functions, a question arises as to how the hierarchy and commitment abilities are formed. The problem is greatly simplified under the punishment assumption introduced above. For, under our assumption of the existence of punishment actions, the only hierarchical position worth having is the first one. Competition for the top spot would plausibly occur in a warlike game to first establish a commitment to punish deviant players. As this battle for dictatorship preceeds our basic game, and has no influence on its solution given the remaining resources and the identity of the victor, we need not model it here. It remains true, however, that the various solutions to our basic game—corresponding to the various possible dictators—do generally affect the identity of the victor in the dictatorship battle. For example, more benevolent players meet with less resistance than their less benevolent competitors and are therefore more likely to wind up as dictator. Also, to reduce the resources devoted to subsequent battles in a life cycle environment, a dictator is likely to train and appoint the next generation's dictator.

C. *A Possible Application to Explaining Broad Features of Observed Institutions*

We can use our special model with its optimality feature to offer an explanation, albeit speculative, of the broad features of observed political-economic institutions without imposing a standard, paradoxical who-guards-the-guards enforcement mechanism and without assuming some sort of inexplicit cooperation among individuals (i.e., social contracts).[2]

To see this let us view the set of individuals $\{1, a(m),\dots, m)$ as a hierarchy representing the individuals in a military chain of command subject to a dictator, individual 1. The military hierarchy rationally establishes reaction functions which ensure the dictator's benchmark set of actions x^1 are carried out. This appears to work well for families and small, tribal societies, which are decidedly hierarchal, stable, command societies (see Manning Nash). Although many tribal societies admit some form of private property and exchange in final consumer goods, this is apparently because of the lack of information on the part of the leader concerning other's preferences combined with the fact that such exchanges do not harm the leader. Free exchange in inputs is a different story. Here the leader's income depends substantially on how inputs are used and here the tribal leader maintains substantial direct control (see M. Nash).

For larger societies, it becomes implausible to assume that the first strategy selector knows what is to him the relevant set of feasible individual actions or the relevant particular actions that are actually undertaken. That is, it is implausible to assume that he knows the capacities and actual performances of his subjects. To get around these information problems and still achieve about the same efficient allocation, the first strategy selector could appoint local leaders who could fairly easily discover the potentials and monitor the actions of the peasants in his territory. When the first strategy selector cannot, in our sense, punish the local leaders, and the latter thereby become about as wealthy as the former, the local leaders can afford the risk of ownership and—because of the obvious incentive value of their being residual claimants given that the first strategy selector cannot easily observe their behavior—they would rationally become landlords over their territories, paying mainly lump sum rents to the overlords. This is, in essence, feudalism. Since a local lord, or vassal, would not differ much with the overlord regarding the welfare of the lord's subjects, little conflict would arise on this issue. Therefore—assuming no technical interdependence between the regions—the vassals would make about the same efficient decisions as the super-informed overlord appearing in our formal model.

Once the overlord acquires punishment power over the local leaders, as occurred a few centuries ago when the invention of gunpowder and related military improvements enabled overlords to readily destroy their feudal fiefdoms (see Richard Bean; Ronald Batchelder and Herman Freudenberger), the wealth of the local leaders is scaled way down as our dictatorial solution takes over. This redistribution means that local leaders can no longer afford the risk of territorial ownership and therefore are paid relatively fixed wages rather than substantial shares of their territorial products. They then have inefficient incentives to extract benefits from the subjects in ways that are not observable to the overlords. To protect the masses from these costly, hidden redistributions to these local leaders, private property could be extended to inputs for the masses with no reduction in theoretical efficiency (see our 1980 paper). To protect individual targets ripe for costly political extortion, the local leaders could be elected by the wealthiest residents of the area. These adjustments are obviously fairly descriptive of the actual institutional transition from feudalism to modern nation statehood. (Alternatively, especially under conditions soon to be specified, a rational overlord might revert to extremely rigid, centrally directed, input controls to avoid the problem of costly expropriation by local leaders that arises when he gains dictatorial power.)

To continue the application, our industrial revolution, fostered by the greater internalization of innovative benefits permitted by the larger governmental units, introduced technologies of easy duplication rather than revolutionizing the quality of the goods that a dictator could receive. Hence, assuming

[2] The weakness of conventional cooperative game theory in describing cooperative behavior is discussed in Section V.

some benevolence on the part of the various dictators, the new dictatorial solutions gave many more goods to the subjects. The much greater wealths of the subjects then made it attractive for efficient dictators to adopt popular democratic voting systems, observed democracies having provided alterations in private property systems that are far superior for the adult voters to those provided by benevolent dictators armed with the best economic theories (for example, see Thompson, 1974; 1979), the only serious drawback of democracy being that the adult voters will inefficiently exploit the children (for example, by overwork and undereducation) when the typical voter is very poor (see Thompson and Wayne Ruhter).

Reinforcement for the above application of our model is found in the essential replication of the observed developmental pattern during the historical period immediately preceeding the decentralizing barbarian invasions that led to our medieval slump. Classical Greek, Etruscan, and Early Roman City States emerged from small, feudal clans once iron smelting made large, destructive, heavy-armed, hoplite armies a practicality during the eighth and seventh centuries B.C. (See, for example, Antony Andrewes or Jacques Heurgon.) Private property in inputs quickly followed in the seventh and sixth centuries B.C., thereby halting a costly expropriation of lower-level citizens by the middle level (see Andrewes; Heurgon). Finally, the greater internalization of innovative returns afforded by the larger-scale governmental units induced a wave of innovations from the sixth century onward that greatly increased the welfare of the lower strata of society and led to the gradual adoption of democratic political structures from the fifth century B.C. onward within the wealthiest, most education-conscious, states (i.e., Athens, the Ionian City States, and Rome (see Andrewes, Heurgon)).

The apparent control that voters have over the benefits of their dictators (military leaders) in democratic countries is, we submit, an illusion. Military leaders, as a group, have both tenure and an absence of significant institutional constraints on the goods and services they can command. The frequency of military takeovers of democratic governments that generate unsatisfactory results from the standpoint of the military in medium-poor countries is evidence for the dominance of the military. It is also further evidence, given the additional fact that wealthier countries typically have popular democracies while poorer countries typically have direct military rule, for the above argument on the rationale for popular democracies. Additional evidence in support of this illusion, together with an argument explaining how the illusion is in the joint interest of the military leadership and the citizens, appears in Thompson (1979, Section 1E).

While most modern analyses of democracy recognize its underlying internal efficiency tendencies in that an allocation cannot be an equilibrium under democracy if there is an alternative allocation in which all voters would be better off, the analyses are also critical of democracy in that it is—within the standard model—unable to achieve a determinate solution. Cyclical majorities, or "voters' paradoxes," arise. They mean either a never-ending series of generally undesirable redistributions and a corresponding drain on society's resources or such severe constraints on the agenda that it is likely that many efficiency-enhancing bills will never be voted upon. But the arguments for voters' "paradoxes" do not arise under truly perfect information! (See Section V.) In fact, very few legislative ballots are secret. It is relatively easy to observe votes and communicate voting reaction functions in legislatures. So the paradoxes appear to exist only within highly inappropriate theoretical models.

We have used the strong efficiency and distributional implications of our model to help predict the occurrence of: (a) feudalism or modern-type statehood; and (b) democracy or authoritarian government. We can also use these implications to predict whether a nation will adopt socialism or capitalism. We have been assuming that the various territories of a country are technically independent. Communication between leaders of the various areas is of no importance under this condition. But suppose that each area has a unique comparative advantage at producing a particular durable input and that these inputs are complementary. If strategic interaction between the leaders of these areas is sufficiently costly that the leaders will not initially (say within an effective legislature) strategically communicate—but not so costly that it would fail to emerge even after certain investments were made—then none of these areas will produce their particular complement and the decentralized nation would be underdeveloped (see Thompson, 1981). In this case, with democratic legislatures ineffective, authoritarian control of investment is required for the nation to reach a developed state. This may explain why resource-rich nations like China and Russia, nations with historically high geographic barriers to interregional interactions and therefore notoriously uncohesive regional leaders (see S. N. Eisenstadt), have employed centralized control of investment in emerging from feudalism to modern nation statehood.

D. *Rationalization for Hierarchies in Nature*

The observed universality of hierarchal social organization among social primates (see Peter Farb), while supportive of our basic assumption, is not directly explainable by the social efficiency of such organization. While would-be dictators generally favor this form, would-be subjects may well be better off without the physical ability to receive or understand the reaction functions of others. Our explanation for the predominance of hierarchal organization is that since most primate evolution has take place in isolated families or small, family-like clans, wherein the members live or die together on the basis of the joint efficiencies of their separate groups, biological evolution has selected against families whose individuals could not receive or communicate reaction functions.

V. Contrast to Other Theories of Conflict Resolution

A. *Cooperative Game Theory*

Cooperative game theory is founded on the assumption that any subset of *n* will form a "blocking coalition," a group of players that interact to preclude a social outcome, whenever an internally feasible alternative would yield a greater payoff. The assumption guarantees each player a payoff at least equal to the minimum of what he can achieve in a one-man coalition. On the additional assumption that any Pareto optimum can be achieved by a coalition of all *n* players, the theory guarantees that any solution must be Pareto optimal. For any Pareto nonoptimal solution would be blocked by an *n*-person coalition. The theory is then devoted to the search for a solution out of the resulting set of "imputations," that is, Pareto optimal points which give each player at least the minimum of what he would receive in a one-man coalition. The standard solution set of recent years, the core, is the set of unblocked outcomes. A chronic problem with this theory is that its solution sets are often empty. Other solution sets, such as von Neumann-Morgenstern's "stable set" and the "bargaining set," are less frequently empty but have the chronic problem of admitting a superabundance of outcomes in their solution sets (see, for example, Guillermo Owen).

We object to cooperative game theory because of its inexplicit communication process and related absence of committed strategies. These weaknesses result in insufficient constraints on the actions of blocking coalitions. This point requires some elaboration.

Blocking coalitions exist in a general form as a by-product of interaction under truly perfect information. For any subset of reaction functions effectively blocks all outcomes which do not simultaneously satisfy these functions; and the players in the subset may be thought of as a blocking coalition. However, in our model, the players may be worse off under their blocking behavior but still engage in it because of their prior commitments. At the same time, these commitments prevent them from participating in subsequent blocking coalitions merely because they would be better off in these coalitions given the subsequently chosen strategies of some other players. To admit such unconstrained coalitional activity would be to deny the original commitments.

Consider, for example, a "majority game," a three-person, zero-sum, game in which, say, a dime and a nickel are to be shared by the three players. If players 1 and 2 each select certain actions implying that they "get together," 2 gets a dime and 1 gets a nickel. If 1 and 3 each select certain actions, where the action is different for 1 than in the former case, then 1 gets a dime and 3 gets a nickel. If 2 and 3 each select new actions implying that they get together, then 3 gets a dime and 2 gets a nickel. Cooperative game theory offers no meaningful solution to this game because, for any distribution of coins, there is a blocking coalition.[3] Under truly perfect information, where the order of strategy selection is, say, 1,2,3, player 1 will adopt the following strategy: "I will get together with 2 if he gets together with me; otherwise, I will perform my part of getting together with 3." Player 2 then selects: "I will perform my part of getting together with 1 regardless of the action of player 3." Player 3 gets nothing no matter what he does. It is easy to verify that there is no other solution. In sharp contrast, under cooperative game theory, 3 would offer to get together with 1, who—being unable to commit himself to a fixed strategy—would be unable to refuse the offer. And we would be off on the never-ending cycle of coalition formation characteristic of existing cooperative game theory.

B. *Voting Theory*

A specification of these acts of getting together enables us to see that voters' paradoxes cannot arise under truly perfect information. Instead of the above world as one with three abstract "meetings," think of the world as one that uses majority rule voting over three possible bills. Correspondingly, let the abstract getting together of players 1 and 2 represent their both supporting the bill that shuts out player 3, the getting together of players 1 and 3 represent 1 and 3's supporting the bill that shuts out 2, and the getting together of 2 and 3 represent their supporting the bill that shuts out 1. It is easy to see that whatever bill becomes law, one of the other bills will defeat it *if* voting follows only the narrow self-interest of the voters. This is the voters' paradox. But the paradox does not arise under the communication of commitments to narrowly irrational but, of course, broadly rational voting strategies. Following our example, player 1 rationally adopts: "I will vote to shut out player 3 if player 2 does, otherwise, I'll vote to shut out player 2." Player 2 then votes to shut out player 3 and player 3 is shut out. While player 3 tempts player 1 with a payoff of 10 if he will vote instead to shut out player 2, player 1 is committed not to so vote. If he were not previously committed to some strategy prior to 3's strategy choice, player 1 would himself be shut out by players 2 and 3.

C. *Supergame Theory*

Supergame theory, like cooperative game theory, is a result of the inability of standard noncooperative game theory to allow for sufficient communication to generally achieve Pareto optimal outcomes. The advantage of supergames relative to cooperative games is the absence of imposed, collective rationality conditions. Supergames are the result of the temporal replication of ordinary two-person games in which strategy sets are expanded to include actions contingent on actions in prior games. The standard supergame strategy, due to R. Duncan Luce and Howard Raiffa, is to play a Pareto optimal action if the other player has played his corresponding Pareto action in the preceeding period; and otherwise play a Nash action for the remainder of the supergame. One type of supergame is the Luce-Raiffa, finite-horizon supergame; the other is the Aumann-Friedman, infinite-horizon supergame.

In the finite-horizon supergame, a problem arises in that it generally pays each player to play a Nash action in the last period. This shortens the supergame by one period; but the same argument applies to the shortened supergame and continues applying up through the first period. Thus, the only solution is the Nash solution. Luce and Raiffa argue that the Nash strategy is dominated by their supergame strategy, where they leave to

[3] While the core and vNM's stable set are empty, the bargaining set contains all possible allocations.

be determined the point at which it pays a player to switch to a Nash action. The problem is that a determination of this point, playing the supergame game as a Nash game, reveals that it always pays to switch just before the other player switches. So the players switch to a Nash solution in the first period. It never pays to play their Luce-Raiffa strategy in a finite supergame (see Anatol Rapoport and A. M. Chammah, and Reinhardt Selten). An alteration of the game, in which one player can communicate a fixed, committed strategy to the other before the latter chooses his strategy will change this result; but it also will make the supergame an unnecessary construct, as we have seen.

An infinite supergame does not have this problem, for there is never a last period in which it pays to switch to a standard Nash solution. However, the standard Nash solution is always a possible solution to the supergame. If one player plays an ordinary Nash strategy, so does the other. So, while the Luce-Raiffa strategies—extended to infinite replication—are a possible solution; so are the ordinary Nash strategies. Furthermore, as Friedman shows, sufficiently high discount rates will assure a Nash solution. Still further is the weakness that the theory does not determine which of the generally many outcomes which are Pareto superior to the Nash Solution will be actually chosen.

But the most obvious weakness of supergame theory is the requirement of an infinite horizon. While it is plausible to assume that individuals behave as if the *world* may last forever, it is implausible to assume that individuals behave as if they, as continually acting individuals capable of continually exhibiting strategies, may last forever.

REFERENCES

Antony Andrewes, *The Greeks*, Norton 1978.

Kenneth J. Arrow, *Social Choice and Individual Values*, 2d ed., New York 1963.

R. J. Aumann, "Acceptable Points in General Cooperative *n*-person Games, Contributions to the Theory of Games, IV," in *Annals of Mathematics Study*, Vol. 40, Princeton 1959, 287–324.

R. W. Batchelder and H. Freudenberger, "A Theory of the Rational Evolution of the Modern Centralized State," discus. paper, Tulane Univ., Oct.1979.

R. N. Bean, "War and the Birth of the Nation State," *J. Econ. Hist.*, Mar. 1973, *33*, 203–21.

S. N. Eisenstadt, "Cultural Orientations, Institutional Entrepreneurs, and Social Change: Comparative Analysis of Traditional Civilizations," *Amer. J. Sociology,* Jan. 1980, *89*, 840–69.

Peter Farb, *Humankind*, Boston 1978, ch. 17.

J. W. Friedman, "A Non-Cooperative Equilibrium for Supergames," *Rev. Econ. Stud.,* Jan. 1971, *38*, 1–12.

Jacques Heurgon, *The Rise of Rome*, Berkeley; Los Angeles 1973.

Nigel Howard, *Paradoxes of Rationality*, Cambridge 1971.

R. Duncan Luce and Howard Raiffa, *Games and Decisions*, New York 1957, ch. 5.

John Nash, "Noncooperative Games," *Annals of Mathematics*, 1951, *54*, 286–95.

Manning Nash, *Primitive and Peasant Economic Systems*, San Francisco: Chandler, 1966, chs. 1–4.

Guillermo Owen, *Game Theory*, Philadelphia 1969.

Anatol Rapoport and A. M. Chammah, *Prisoner's Dilemma*, Ann Arbor 1965, 26–29.

R. Rosenthal, "Induced Outcomes in Cooperative Normal Form Games," discus. paper 178, Center for Mathematical Studies in Economics and Management Science, Northwestern Univ., November 1975.

Thomas C. Schelling, *The Strategy of Conflict,* Cambridge 1963.

R. Selten, "The Chain-Store Paradox," working paper no. 18, Institute of Mathematical Economics, Univ. Bielefeld 1974.

E. A. Thompson, "Taxation and National Defense," *J. Polit. Econ.*, July/Aug. 1974, *82*, 755–83.

______, "An Economic Basis for the 'National Defense Argument' for Protecting Certain Industries," *J. Polit. Econ.*, Feb. 1979, *87*, 1–36.

______, "The Value of Information in Non-Conflict Situations," working paper, Univ. California-Los Angeles 1981.

______ **and Roger L. Faith,** "A Model of Non-

Competitive Interdependence and Anti-Monopoly Law," working paper 143, Univ. California-Los Angeles, Jan. 1979.

______ and ______, "Social Interaction under Truly Perfect Information," *J. Math. Sociol.*, Part IV, Nov. 1980, 7, 181–97.

______ and **Wayne E. Ruhter,** "Parental Malincentives and Social Legislation," working paper 141, Univ. California-Los Angeles 1981.

John von Neumann and Oscar Morgenstern, *Theory of Games and Economic Behavior*, 3d ed., Princeton 1953.

R. Wilson, "The Game Theoretic Structure of Arrow's General Possibility Theorem," *J. Econ. Theory*, Aug. 1972, *5*, 14–20.

Optimal Auctions

By JOHN G. RILEY AND WILLIAM F. SAMUELSON*

In the two decades since the seminal paper by William Vickrey, literature on the theory of auctions has developed at a rapid though uneven pace.[1] Much of this literature is fragmentary, varies widely in scope, and is not easily accessible to economists. As a result, the implications of different auction rules in various settings remain relatively unknown. This paper provides a systematic examination of alternative forms of auctions. In so doing it presents a general characterization of the implications for resource allocation of different auction designs within the model originally proposed by Vickrey.

The auction model is a useful description of "thin markets" characterized by a fundamental asymmetry of market position. While the standard model of perfect competition posits buyers and sellers sufficiently numerous that no economic agent has any degree of market power, the bare bones of the auction model involves competition on only one side of the market. In this setting a single seller of an indivisible good faces a number (n) of potential buyers. Competition among the (possibly small number of) buyers takes place according to a well-defined set of auction rules calling for the submission of price offers from the buyers. Most commonly, the choice of auction method employed rests with the monopolistic seller.

These brief observations suggest two natural questions for analysis: First, what form does the competition among the few buyers take under the most common auction procedures? In turn, how is a sale price determined? Second, by what means can the seller best exploit his monopoly position? For example, would it be more profitable for the seller to require payment not only by the high bidder, but also by those with lower ranked bids?[2]

As one might expect, any change in the rules of the auction results in different bidding strategies on the part of the buyers. In particular, if the auction rules posit a minimum payment by one or more of the bidders (determined by rank), those with sufficiently low reservation values will be discouraged from entering a bid. Our analysis will demonstrate that, in a risk-neutral setting, it is the reservation value below which a buyer opts to remain out of the auction which is crucial. To be precise, if the lowest reservation value for which it is worthwhile bidding is the same for two different auction rules, then the expected return to the seller is also the same.

Throughout the paper we shall retain the following basic assumption.

a) A single seller with reservation value v_0 faces n potential buyers, where buyer i holds reservation value v_i, $i=1,\ldots,n$.

b) The reservation values of the parties are independent and identically distributed, drawn from the common distribution $F(v)$ with $F(\underline{v})=0$, $F(\bar{v})=1$ and $F(v)$ strictly increasing and differentiable over the interval $[\underline{v},\bar{v}]$. We will refer to this as the *IID* assumption.

The *IID* assumption was first presented by Vickrey, and has been frequently employed

*Professor of economics, University of California-Los Angeles, and assistant professor, School of Management, Boston University, respectively. Riley's research was supported by National Science Foundation grant SOC79-07573. Helpful discussions with Howard Raiffa, Jerry Green, Eric Maskin, and Paul Samuelson, and the valuable comments of a referee are gratefully acknowledged.

[1] A current bibliography by Robert Stark and Michael Rothkopf lists nearly 500 papers written over this period. For a recent survey of this literature, see Richard Engelbrecht-Wiggans.

[2] Vickrey's comparison of the open "ascending bid" auction and the sealed "high bid" auction has been generalized in unpublished dissertations by Armando Ortega-Reichert, Gerard R. Butters, and William Samuelson. Milton Harris and Artur Raviv (1979) provide the first discussion of optimal auction design. Employing very different methods they consider the special case in which each buyer has flat (uniform) prior probabilistic beliefs about the amount others are willing to pay.

in the bidding literature. In practical terms, each party is uncertain about the others' reservation values, believing that each individual decides the maximum amount he is willing to pay independently of the others. In addition, the parties share common priors with respect to the possible reservation values of each individual. With the *IID* assumption, the bidding procedures we outline below belong to the class of games of incomplete information first formulated by John Harsanyi.

The paper is organized as follows: First we present our central result demonstrating that expected seller revenue from quite different auctions can be very easily compared. As an immediate implication, the equivalence of the "English" or "ascending bid" auction and the "Dutch" or "high bid" auction is established. More important, it is shown that, for a broad family of auction rules, expected seller revenue is maximized using either of the two common auctions if the seller announces that he will not accept bids below some appropriately chosen minimum or "reserve" price. Surprisingly, this reserve price is independent of the number of buyers and is always strictly greater than the seller's personal value of the object. In Section II several alternative auction rules are described in detail and their implications for the seller are compared. Finally, in Section III, the two commonly used auctions are once again compared under the assumption that the buyers are risk averse rather than risk neutral. It is shown in this setting that the English auction is dominated by the sealed high bid auction, and that the optimal reserve price in the latter is a declining function of the degree of buyer risk aversion.

I. Comparison of Alternative Auction Rules

Before characterizing a broad family of alternative auction rules, a few remarks about the English or ascending bid auction will be helpful.

In auctions of antiques, estate objects, and works of art, the good is awarded to the buyer who makes the final and highest bid. The buyer placing the highest valuation on the good therefore pays approximately the maximum of the reservation values of the other $n-1$ buyers. As Vickrey noted, this is equivalent to a sealed bid auction in which each buyer submits a bid and the high bidder pays the *second* highest rather than the highest bid.[3] To see this, suppose the ith buyer considers shading his bid, b_i, below his reservation value v_i. If the largest of all the other bids, b_*, exceeds v_i, another buyer is the high bidder so that buyer i's gain remains zero. If $b_* < b_i$, buyer i remains the high bidder and continues to gain $v_i - b_*$. However, if $b_i < b_* < v_i$, the shading yields a zero gain, whereas without shading the gain is $v_i - b_*$. A parallel argument establishes that there is no advantage in making a bid, b_i, greater than v_i. The optimal strategy of each buyer is therefore to submit his reservation value. It follows that, just as in the English auction, the high bidder ends up paying the second highest reservation value. This equivalence greatly simplifies the comparison between the English and sealed high bid auctions[4] since it implies that we need only compare sealed bid auctions.

In the high and second bid auctions, only the winner makes a payment to the seller. However, there is an infinity of auction rules involving payment by more than one bidder. For example, all buyers might be charged a fixed entry fee. Alternatively, losers might be required to pay some fraction of their bids. A third possibility, discussed in Section II, is that the seller might attempt to encourage higher bids by offering to return some of the money paid by the winner to each of the losers, the size of the rebate depending on a loser's bid.

Each of these alternatives is an example of an auction with the following properties. First, a buyer can make any bid above some minimum "reserve" price announced by the seller. Second, the buyer making the highest bid is awarded the object. Third, the auction rules are anonymous: each buyer is treated

[3] This type of auction is sometimes referred to as a Vickrey auction.

[4] The sealed high bid auction also has its open auction equivalent. In this Dutch auction, the sale price is initially set at a high level and is then lowered until a bid is made.

alike. Fourth, there is a common equilibrium bidding strategy in which each buyer makes a bid b_i, which is a strictly increasing function of his reservation value v_i, i.e.,

$$b_i=b(v_i) \quad i=1,\ldots,n \tag{1}$$

Throughout this section we shall consider the family $\mathcal{A}$ of auction rules for which these four assumptions are satisfied.[5] We begin with the main result.

PROPOSITION 1: *Suppose the IID assumption holds and all buyers are risk neutral. The common equilibrium bidding strategy for any member of the family $\mathcal{A}$ of auction rules yields an expected revenue to the seller of*

$$n\int_{v_*}^{\bar{v}}(vF'(v)+F(v)-1)F(v)^{n-1}dv$$

where v_* is the reservation value below which it is unprofitable to submit a bid.

Proposition 1 is important because it tells us that the expected revenue of the seller from quite different auctions can be compared simply by determining the lowest reservation value, v_*, for which it is worthwhile bidding. We begin the proof by examining the behavior of a single buyer. The expected return to making a bid can be expressed as follows.

$$\left\{\begin{matrix}\text{expected}\\ \text{buyer}\\ \text{gain}\end{matrix}\right\}=\left\{\begin{matrix}\text{reservation}\\ \text{value}\end{matrix}\right\}\left\{\begin{matrix}\text{probability}\\ \text{of}\\ \text{winning}\end{matrix}\right\}-\left\{\begin{matrix}\text{expected}\\ \text{payment}\end{matrix}\right\} \tag{2}$$

Below we obtain simple expressions for both the probability of winning and the expected buyer gain. Then, from (2), we are able to derive the expected payment of a typical buyer. Given the symmetry of the auction rule, expected seller revenue is just n times this expected payment.

With buyers behaving noncooperatively, a common strategy, $b_i=b(v_i)$, is an equilibrium strategy if, when adopted by all buyers but one, the latter's best response is to adopt it also. Without loss of generality, we may suppose that the buyer considering an alternative bidding strategy is buyer 1. With all other buyers bidding according to $b(v)$, buyer 1, if he bids at all, will wish to bid in the range of this function. Hence, we can write any bid as $b_1=b(x)$ and view buyer 1 as choosing x. It follows that $b(v)$ is an equilibrium bidding strategy if buyer 1 can do no better than choose $x=v_1$, and so bids $b(v_1)$.

To examine the optimal choice of buyer 1, we begin by assuming his bid is $b(x)$, and then ask what restrictions are implied by the requirement that his optimal bid is $b(v_1)$. Any auction rule must specify the amount he must pay, p, given his own bid $b_1=b(x)$ and those by the other $n-1$ buyers, i.e.,

$$\begin{aligned}p&=p(b_1,b_2,\ldots,b_n)\\ &=p(b(x),b(v_2),\ldots,b(v_n))\end{aligned}$$

We may therefore write the expected payment by buyer 1, given a bid of $b_1=b(x)$ as

$$P(x)=\underset{v_2,\ldots,v_n}{E}\,p(b(x),b(v_2),\ldots,b(v_n)) \tag{3}$$

Also, the bid of $b(x)$ is the winning bid if and only if all other buyers have made lower bids. By assumption the equilibrium bid function is strictly increasing in v, therefore buyer 1 wins if all other valuations are less than x. Since the probability buyer j has a reservation value less than x is $F(x)$, buyer 1 wins with probability $F^{n-1}(x)$. Combining this last result with (2) and (3), the expected gain to buyer 1, if he chooses to enter the auction, can be expressed as

$$\Pi(x,v_1)=v_1F^{n-1}(x)-P(x) \tag{4}$$

[5] After deriving Proposition 1, we became aware of a paper by Roger Myerson which uses a much more technically demanding approach to examine expected seller revenue in an even broader class of auctions. Generalizing our approach, Eric Maskin and Riley (1980a) have shown that Myerson's results imply there there are circumstances in which expected seller revenue can be increased by prohibiting bids over certain ranges.

For $b(v)$ to be the equilibrium bidding strategy, buyer 1's optimal choice must be to select $x=v_1$ and bid $b(v_1)$. Then buyer 1's maximized expected gain is $\Pi(v_1, v_1)$ and the following first-order condition must be satisfied.[6]

$$(5) \quad \frac{\partial \Pi}{\partial x}(x, v_1) = v_1 \frac{d}{dx} F^{n-1}(x) - P'(x) = 0 \quad \text{at } x = v_1$$

This must hold for all reservation values exceeding v_*, the reservation value for which a buyer is indifferent between submitting the bid, $b(v_*)$, and not entering the auction. That is, (5) holds for all $v_1 \geqslant v_*$, where v_*, satisfies

$$(6) \quad \Pi(v_*, v_*) = v_* F^{n-1}(v_*) - P(v_*) = 0$$

Setting $x=v_1$ in (5), it follows that the equilibrium expected payment by buyer 1 must satisfy the differential equation

$$(7) \quad P'(v_1) = v_1 \frac{d}{dv_1} F^{n-1}(v_1) \qquad v_1 \geqslant v_*$$

Integrating and making use of the boundary condition, (6), buyer 1's expected payment is therefore

$$(8a) \quad P(v_1) = v_* F^{n-1}(v_*) + \int_{v_*}^{v_1} x \, dF^{n-1}(x) \qquad v_1 \geqslant v_*$$

Integrating the second term by parts, this can be rewritten more conveniently as

$$(8b) \quad P(v_1) = v_1 F^{n-1}(v_1) - \int_{v_*}^{v_1} F^{n-1}(x) \, dx \qquad v_1 \geqslant v_*$$

The final step is to consider the auction from the seller's viewpoint. As far as the seller is concerned, v_1 and hence the expected payment, $P(v_1)$, is a random variable. The seller's expected revenue from buyer 1 is therefore the expectation of $P(v_1)$. Since the seller knows that v_1 has distribution $F(v_1)$ his expected revenue is

$$\bar{p}^1 = \int_{v_*}^{\bar{v}} P(v_1) F'(v_1) dv_1$$

Substituting for $P(v_1)$ from (8b) and integrating by parts, the expected revenue from buyer 1 can be rewritten as follows:

$$(9) \quad \bar{p}^1 = \int_{v_*}^{\bar{v}} \left[vF'(v) + F(v) - 1 \right] F^{n-1}(v) dv$$

Given the equal treatment of all n buyers, expected seller revenue is just n times the expected revenue from buyer 1 and the proposition is proved.

One of the striking features of our derivation is that nowhere is there explicit reference to the equilibrium bidding strategy $b(v)$. However, under any particular auction rule this is readily derived. The key to such derivation is (8), the expression for expected payment, $P(v)$, of a buyer with reservation value v.

For example, in the high bid auction, suppose the seller announces a reserve price b_0. Any buyer with a reservation value $v > b_0$ has an incentive to enter, i.e., $v_* = b_0$. Since a buyer pays if and only if he is the high bidder, his expected payment is

$$(10) \quad P(v) = Prob\,\{b(v) \text{ is high bid}\}\, b(v)$$

But $b(v)$ is the high bid if and only if all other buyers have lower reservation values. Then $Prob\,\{b(v) \text{ is high bid}\} = F^{n-1}(v)$ and, from (10), $b(v) = P(v)/F^{n-1}(v)$. Substituting for $P(v)$ from (8b), we therefore have the following additional result.

PROPOSITION 2: *Suppose the IID assumption holds, all buyers are risk neutral and the seller announces a reserve price b_0. In the high*

[6] From (4) and (5) we also have

$$\frac{\partial \Pi}{\partial x}(x, v_1) = (v_1 - x)(n-1) F^{n-2}(x) F'(x)$$

Therefore $\Pi(x, v_1)$ is increasing in x for $x < v_1$ and decreasing for $x > v_1$ and the first-order condition yields the global maximum for buyer 1.

bid auction, the equilibrium bidding strategy of a typical buyer with reservation value $v \geqslant b_0$ is

$$b(v)=v-\int_{b_0}^{v} F^{n-1}(x)dx/F^{n-1}(v)$$

Proposition 2 indicates the degree to which a buyer will "shade" his bid, $b(v)$, below his reservation value v in the high bid auction. It is a straightforward matter to confirm that $b(v)$ is strictly increasing in v. Therefore the high bid auction is a member of the family of auctions described by Proposition 1. Certainly the second bid auction is in this family since anyone entering will bid his reservation value. In both auctions, a buyer will enter if and only if his reservation value exceeds b_0. Then in both cases $v_*=b_0$ and, from Proposition 1, expected seller revenue is the same.

We now demonstrate that these two common auction rules are optimal for the seller, given the appropriate choice of a reserve price. For any auction rule there is some implied minimum reservation value v_* below which buyers will choose not to bid. Then there is a probability of $F^n(v_*)$ that all n buyers will decide not to submit a bid. In this case the seller's gain is his own personal valuation v_0. Then, from (9), the total expected return to the seller is

$$v_0F^n(v_*)+n\int_{v_*}^{\bar{v}}(vF'(v)+F(v)-1)F^{n-1}(v)dv \tag{11}$$

It follows that any two auctions in the family $\mathscr{A}$, for which v_* is the same, yield the same expected gain to the seller.[7] Moreover, differentiating with respect to v_* the expected gain of the seller is maximized for some v_* satisfying the condition[8]

$$n\left[v_0F'(v_*)-v_*F'(v_*)-F(v_*)+1\right]F^{n-1}(v_*)=0 \tag{12}$$

We therefore have the following further result.

PROPOSITION 3: *If the IID assumption holds and buyers are risk neutral, the members of the family $\mathscr{A}$ of auction rules, which maximize the expected gain of the seller are those for which the reservation value v_*, below which it is not worthwhile bidding, satisfies*

$$v_*=v_0+1-F(v_*)/F'(v_*)$$

independent of the number of buyers.

An immediate implication of Proposition 3 is that the high and second bid auctions with reserve price $b_0=v_*$ are both optimal in the family of auctions $\mathscr{A}$. Note also that v_* exceeds v_0: the seller announces a reserve price strictly greater than his personal valuation.

To gain an understanding of this strong result, it is helpful to consider a second bid auction in which there are two buyers and to examine the implications of introducing a reserve price slightly higher than v_0; i.e., $v_*=v_0+\delta$ where δ is small. Since each buyer's dominant strategy is to bid his reservation value, the expected gain to the seller is affected (i) if both valuations lie between v_0 and v_*, and (ii) if one valuation lies between v_0 and v_* and the other exceeds v_*. In the first case the seller retains the item, which he values at v_0, rather than selling it at some price between v_0 and v_*. His loss is therefore of order δ. Since this outcome occurs with probability $(F(v_*)-F(v_0))^2 \approx (F'(v_0))^2\delta^2$, the expected loss is of order δ^3. In the second case the seller receives a payment of v_* rather than some price between v_0 and v_* hence has a gain of order δ. Since this outcome occurs with probability $2(F(v_*)-F(v_0))(1-F(v_*)\approx 2F'(v_0)(1-F(v_0)\delta$, the expected gain is of order δ^2. Therefore, for sufficiently small δ, the gain to raising the reserve price above the seller's personal valuation outweighs the cost.

While we have focused on the reserve price because of its common usage, there are many different ways in which to discourage the appropriate subset of buyers from participating in the bidding. Suppose, for example,

[7] Moreover, any two auctions for which v_* is the same yield the same expected gain to buyer i conditional on v_i. This result follows directly from equations (4) and (8b).

[8] Expression (12) will, in general, have multiple roots. If this is the case, it is necessary to evaluate the expected return, (11), at each root to determine the global maximum.

that the seller announces a fixed entry fee c. For all buyers with valuations less than some number v_c, it will be optimal to remain out of the auction. Consider a buyer with the borderline reservation value v_c. In the second bid auction he enters, and, since the entry fee is now sunk, bids his true value v_c. He wins if and only if there are no other bidders, in which case there is no additional payment. Since this occurs with probability $F(v_c)^{n-1}$ his expected profit is

$$v_c F(v_c)^{n-1} - c \tag{13}$$

But for v_c to be the borderline reservation value, the expected profit must be zero. The seller then chooses an entry fee c_* satisfying

$$c_* = v_* F(v_*)^{n-1} \tag{14}$$

A similar argument holds for the high bid auction. If a buyer has the borderline reservation value v_c, he wins if and only if there are no bidders. The optimal bid in such circumstances is zero, therefore the expected profit is again given by (13) and the optimal entry fee by (14).

Our general results are also helpful in analyzing the expected payoff to multiple rounds of bidding. Suppose, for concreteness, that the seller is using a high bid auction. If he can convince buyers that there will only be a single round with optimal reserve price v_*, there will be a chance that no bids will be submitted. Since v_* exceeds the seller's reservation value v_0, the seller, after the fact, has an incentive to lower his reserve price and to call for a second round of bids. However, if buyers are not fooled, they will adopt first-round strategies in the expectation of a possible second round. In particular, all those with reservation values $v_j > b_{00}$, the reserve price in the second round, will plan to enter the two round auction. Then, from Proposition 3 the seller's expected gain is lower since the entry value is no longer optimal.

A final point concerns the decision of the seller whether or not to announce a reserve price. In the second bid auction, the strategy of bidding one's reservation value is a dominant strategy. Therefore the seller cannot influence bids by concealing his reserve price. It follows that the optimal silent reserve price is the same as the optimal announced reserve price and that expected seller revenue is identical.

The argument is more complex in the case of the high bid auction, but once again it can be shown that there is no advantage in using a silent reserve price. The proof, which involves a straightforward extension of Proposition 1, is provided in our earlier paper.

II. Alternative Auctions

To illustrate the results of Section I, we now compare some specific auctions under the simplifying assumptions that there are only two buyers, and that reservation values are uniformly distributed on the unit interval ($F(v)=v$, for $v \in [0,1]$). We assume also that the object for sale has no value to the seller, $v_0 = 0$. First we indicate the gains to employing an optimal reserve price in the high bid auction. We then present an unusual pair of auction designs which happen to belong to the class of optimal auctions.[9] In contrast, a third example shows that a seemingly natural (and commonly employed) auction procedure is suboptimal.

Under our simplifying assumptions, it follows from Proposition 3 that it is optimal for the seller to design the auction so that only those with reservation values exceeding $v_* = 1/2$ find it worthwhile bidding. Then, in the high bid auction it is optimal for the seller to announce a minimum or reserve price $b_0 = 1/2$. Appealing to Proposition 2, this, in turn, implies that the equilibrium bid of buyer i, with reservation value $v_i \geqslant 1/2$, is $b(v_i) = v_i/2 + 1/8v_i$. By contrast, if the seller always sells the good (by setting the reserve price $b_0 = 0$) buyer i's bid becomes $\tilde{b}(v_i) = v_i/2$. Either by direct computation or by appealing to Proposition 1, it can be confirmed that expected seller revenue is $5/12$ with $b_0 = 1/2$ and $1/3$ with $b_0 = 0$. Thus the optimal re-

[9] The interested reader is referred to our earlier paper, where it is shown that the auctions of examples 1 and 2 belong to the class of optimal auctions for arbitrary $F(v)$ and n.

serve price strategy results in a 25 percent increase in expected revenue.

We now consider three quite different auction rules.

Example 1: *Sad Loser Auction*. Suppose there are just two buyers and the seller announces the following auction rules.

(i) Each buyer paying an entry fee c is eligible to submit as his bid any positive real number.[10]

(ii) The high bidder receives the good but retains his bid.

(iii) The lower bidder (if there is one) loses his bid.

It is tempting to conjecture that there is no equilibrium bidding strategy for this set of rules. However, not only is this incorrect, but the equilibrium bidding strategy is readily derived. Under rules (i)–(iii), the expected gain of the typical buyer is

$$\Pi(x, v_i)=v_iF(x)-b(x)(1-F(x))$$

For $b(v)$ to be the equilibrium bidding strategy, this gain is maximized by setting $x=v_i$ so that the buyer's expected payment is $b(v_i)(1-F(v_i))$. Then if $F(v)=v$ and $c=1/4$, it can be confirmed from equation (14) that $v_*=1/2$, and from equation (8b) that

$$b(v_i)=\frac{(v_i^2-1/4)}{2(1-v_i)} \qquad \text{for } v_i \geqslant \frac{1}{2}$$

Thus, as in the optimal high bid auction, any buyer with reservation value less than 1/2 remains out of the auction. The bids of those with higher reservation values are strictly increasing in v and increase without bound as v approaches 1! Nevertheless, it is easy to confirm that expected seller revenue under this scheme matches that of the high bid and second bid auction, cum optimal reserve price.

Under the high bid and second bid auctions, only the recipient of the good gains. In contrast, the following auction distributes a positive return to all participants.

Example 2: *Santa Claus Auction*. Suppose there are just two buyers, and the seller announces the following auction rules.

(i) A buyer who submits a bid $b \geqslant v_*$ receives from the seller an amount $S(b)=\int_{v_*}^{b} F(v)dv$.

(ii) The high bidder obtains the good for his bid price so that his net payment is $b-S(b)$.

One can confirm that the equilibrium strategy of each buyer is to bid his reservation value. Suppose that the second buyer bids $b_2=v_2$. Then if buyer 1 bids b_1, his expected profit is given by

$$\begin{aligned}Pr\{b_1 \text{ is high bid}\}(v_1-b_1)+S(b_1)\\ =F(b_1)(v_1-b_1)+S(b_1)\end{aligned}$$

It is straightforward to check that this expression is maximized at $b_1=v_1$.

In this auction the seller's expected net revenue is the expected value of the higher of the two bids less the seller's expected payments. With $v_*=1/2$, $S(b)=b^2/2-1/8$. Moreover, each buyer bids his reservation value; therefore the seller's expected gross receipts and payments are easily computed. Once again it can be confirmed that expected net revenue is 5/12, exactly the sum the seller can expect from the high bid auction.

Since the implication of Proposition 1 is that many seemingly different auction techniques lead to the same ultimate results, it is important to illustrate the range of exceptions.

Example 3: *Matching Auction*. Suppose there are just two buyers and the seller employs the following auction rules.

(i) There is a single round of bidding. Buyer 1 is given the opportunity to quote a price $b_1 \geqslant v_*$.

(ii) If buyer 1 makes a bid, buyer 2 can match it, if he chooses, obtaining the good for this price. If buyer 1 makes no bid, buyer 2 can obtain the good at price v_* if he chooses.[11]

[10] This rules out bids such as "infinity" or "one more than my opponent."

[11] For an analysis of the matching auction when m rounds of bidding are permitted, see our earlier paper.

Though this auction procedure is quite common (for example, in house sales, a renter occupant is frequently given the right to match the offer of any potential buyer), it is inefficient from the point of view of the seller. In fact, in some circumstances it permits a buyer who values the item less highly than his opponent to obtain the good. Thus, it may produce an allocation of the good that is inefficient *ex post*.

Suppose that $F(v)=v$ and $v_*=1/2$. The strategy of buyer 2 is straightforward. He matches b, if and only if $v_2 \geqslant b$. If buyer 1 does not open the bidding, buyer 2 bids $1/2$ for the good if $v_2 \geqslant 1/2$. Anticipating the behavior of buyer 2, buyer 1 bids $b_1 \geqslant 1/2$ to maximize

$$(v_1 - b_1)\, Prob\{\text{buyer 2 chooses not to match}\}$$

Buyer 2 will not match if his reservation value is less than b_1, that is, he will not match with probability b_1. Then buyer 1 chooses $b_1 \geqslant 1/2$ to maximize his expected gain $(v_1 - b_1)b_1$. Since this expression is decreasing in b_1 for all $b_1 > 1/2$, buyer 1's optimal strategy is to bid

$$b_1(v)=\begin{cases} 0 & v_1<1/2 \\ 1/2 & v_1 \geqslant 1/2 \end{cases}$$

Consequently, whenever $1/2 < v_2 < v_1$, the object is awarded to buyer 2 who values it less highly than buyer 1. The expected revenue of the seller for this example is $3/8$, a reduction of 10 percent relative to the high and second bid auctions.

III. Buyer Risk Aversion

When potential buyers are risk averse, the fundamental equivalence result outlined in Section I is no longer valid.[12] Retaining the assumption of buyer symmetry, it is shown in the Appendix that the high bid auction dominates the second bid auction under buyer risk aversion.

PROPOSITION 4: *Suppose assumption IID holds and all buyers share a common utility function displaying risk aversion. Then* (i) *In the second bid auction, bidders continue to bid their reservation values, that is,* $b_i = v_i$. (ii) *In the high bid auction, as bidders become more risk averse, they make uniformly higher bids.* (iii) *Consequently, the seller enjoys a greater expected profit under the high bid auction than under the second bid auction.*

It is evident that the introduction of risk aversion does not affect the strategy dominance of bidding one's true reservation value in a second bid auction, hence part (i). Part (iii) follows directly from part (ii) which is proved in the Appendix.

The intuition behind these results is that with risk aversion the marginal increment in wealth associated with a successful, slightly lower bid is weighted less heavily than the possible loss $(v_i - b_i)$ if, as a result of lowering the bid, the buyer is no longer the high bidder. This leads risk-averse bidders always to shade their bids less than risk-neutral bidders.

Under risk aversion, the general equivalence result obtained in Proposition 1 no longer holds. For instance, an auction employing a seller reserve price will not, in general, be equivalent to one that specifies a buyer entry fee—even when the same reservation value v_*, below which it is not worth bidding, is implied. Still it is natural to explore the effect that buyer risk aversion has on the optimal seller reserve price in the high bid auction. The following result is derived in the Appendix.

PROPOSITION 5: *Suppose assumption IID holds and all buyers share a common cardinal utility function. Then, in the high bid auction, the optimal seller reserve price is a declining function of the degree of risk aversion.*

The proposition is intuitively plausible in view of the fact that as buyers become risk

[12] Other authors have also considered the effects of risk aversion on bidding. Butters derives Propositions 4 and 5 for the special case in which buyers exhibit constant relative risk aversion. Charles Holt examines the effects of risk aversion in the closely related problem of bidding on incentive contracts. Steven Matthews compares high bid and second bid auctions when seller and buyers are risk averse.

averse in the extreme, the amount by which they will shade their reservation values approaches zero, $b(v_i) \to v_i$. Naturally, the seller can do no better than to announce his personal valuation as his reserve price, $b_0 = v_0$. To quote a higher price cannot "push up" buyer offers and risks the loss of beneficial sales. Of course when $b_0 = v_0$ and $b_i = v_i$, the high bid auction is also efficient *ex post*.

IV. Concluding Remarks

While a general result concerning the design of optimal auctions under uncertainty has been presented, it is important to point out the limitations and special assumptions of the present model. We have assumed that:

(a) A single indivisible good is to be sold to the highest bidder.

(b) The greater a bidder's reservation value the more he will bid for the good.

(c) Buyer roles are symmetrical (i.e., buyer values are drawn from a common distribution) and each buyer is risk neutral.

(d) Buyer values are independent.

Additional difficulties are raised when multiple goods are auctioned or when a divisible good must be allocated. Unless buyer valuations are additive and income independent, auctioning the goods in sequence will be inefficient (*ex post* and *ex ante*). When multiple goods are auctioned, each buyer should logically submit a bid for each subset of goods. Roughly speaking, the seller will allocate goods to maximize revenue under one of a number of auction schemes. In the case of a divisible good, each buyer will submit a "demand schedule" indicating the price he is willing to pay for any given quantity of the good. The seller must formulate an auction rule which specifies the allocation of the good and appropriate payment of buyers. In either instance the determination of optimal auctions for these more general environments lies beyond the bounds of the present analysis.[13]

Given assumption (b) it follows that the family of auctions considered are those in which the good is sold to the buyer with the highest reservation value v, if the good is sold at all. Under only moderate restrictions on the form of the distribution $F(v)$, it can be shown that it is never optimal to utilize a rule in which the winner might be someone other than the buyer with the highest v. However, when these restrictions are not satisfied, a stochastic auction is optimal. In such an auction a lottery is employed to allocate the good when buyer reservation values fall in specified ranges.[14]

Dropping the assumption of buyer symmetry also causes complications in the analysis. The derivation of the class of optimal auctions relied explicitly on the existence of a *common* equilibrium bidding strategy. Without this, these propositions no longer hold. The asymmetric model, though far more complex, is nevertheless amenable to the basic approach developed herein. Suppose the reservation prices of the buyers are drawn from the independent distributions, $F_1, F_2, \ldots, F_n$. Some partial results from this setting suggest a basic conclusion. An optimal auction extends the asymmetry of the buyer roles to the allocation rule itself. The assignment of the good and the appropriate buyer payment will depend not only on the list of offers, but also on the identities of the buyers who submit the bids. In short, an optimal auction under asymmetric conditions violates the principle of buyer anonymity.

As pointed out earlier, the assumption of risk neutrality is crucial to our general equivalence result. Given risk neutrality, the seller can do no better than to employ the second bid auction with an optimal reserve price. In this auction, buyers will have no difficulty formulating an optimal bidding strategy. Nor need they know the form of the distribution function $F(v)$. Against any distribution of opponents' bids, each buyer's dominant strategy is to bid his reservation value. The clear advantage of the second bid auction is that it economizes on the information each buyer requires to bid optimally. Furthermore, Proposition 3 indicates that the seller

[13] Harris and Raviv (1981) and Maskin and Riley (1980b) analyze optimal auctions for different classes of demand curves.

[14] For a presentation of the more general framework from which the optimal stochastic auction can be derived see Myerson or Maskin and Riley (1980a).

can formulate an optimal reserve price policy without knowledge of the number of buyers who might enter the auction.[15]

Finally, one must consider the appropriateness of the model's most basic assumption, value independence. The analysis has assumed that each buyer is informed of his own reservation price and, more important, that this price conveys no information about any other buyer's value. A different auction model has been applied to bidding for offshore oil leases.[16] Here, a tract being auctioned is assumed to have a common value for all parties. The tract value is unknown, though buyers may possess (differing) sample information allowing inferences about this value. In this setting, each buyer must determine a strategy for acquiring information concerning the value of the tract and for submitting a bid based on a correct estimate of this value. These features have a direct influence on the determination of an optimal auction and raise additional policy issues. (Should the seller maintain a stake in an awarded tract for the purpose of risk sharing? Should the seller undertake measures to facilitate information acquisition or to allow information pooling?)[17]

In most real world settings, we would expect that a good's economic value to a potential buyer consists of two parts—a value element which is common to all market participants and one which is buyer specific. Shell's recent $3.6 billion purchase of Belridge Oil—the most expensive in *U.S.* history—is a dramatic example. Belridge was sold by closed sealed bid auction in which twenty-odd prospective buyers participated. Differences in bids presumably reflected (i) differences in beliefs about the value of Belridge's oil holdings (differences that might have been dissipated through pooling of information) and (ii) differences in the extent to which Belridge's operations complemented bidders' other activities. It is easy to imagine, though not to solve, a hybrid model specifying both dependent and independent components of buyer reservation values. A formal analysis of optimal auction design in this more general environment remains to be undertaken.

Appendix

PROPOSITION 4 (ii): *Suppose assumption IID holds and all buyers share a common utility function displaying risk aversion. Then in the high bid auction, as bidders become more risk averse, they make uniformly higher bids.*

PROOF:

Let $b(v)$ be the common equilibrium strategy of n risk averse buyers, each of whom has the same von Neumann-Morgenstern utility function $u(x)$. We assume that $u(x)$ is a strictly increasing, concave function of x and normalize so that $u(0)=0$. With all other buyers using the equilibrium bidding strategy and buyer j bidding $b(x)$, j's expected utility is

$$F^{n-1}(x)u(v_j-b(x)) \tag{A1}$$

For $b(x)$ to be the equilibrium strategy, (A1) must have its maximum at $x=v_j$. Differentiating with respect to x and setting the derivative equal to zero at $x=v_j$, we have the

[15] The largest auction houses (for example, Sotheby Park Bernet, Inc. and Christie's) employ the English auction (combining its open bid and sealed bid forms) to sell rare and valuable items (art, antiques, and jewelry). A buyer can bid personally for an item on the day of the auction or can submit a prior written offer, designating a representative from the auction house to bid on his behalf. This same procedure establishes a silent seller reserve price, since a house representative is instructed to buy back the good if the sale price is insufficient. It is a common observation that the competitive features of the open ascending auction serve to elevate buyer offers (above their prior values). This implies that the open ascending auction enjoys a practical advantage over the sealed bid version. The "mixed" auction allows written bids in order to promote the greatest possible participation while maintaining the "uplifting" features of the open ascending auction.

[16] See, for example, Robert Wilson (1975) and Matthew Oren and Albert Williams. In this model buyers begin with common prior beliefs about the value of a resource but have different *posterior* beliefs as a result of independent sampling. For discussion of auctions in which buyers have different prior beliefs, see Wilson (1967).

[17] For a discussion of the incentives for the seller to make information public, see Paul Milgrom and Robert Weber.

necessary condition

$$(n-1)F^{n-2}(v_j)F'(v_j)u(v_j-b(v_j)) - F^{n-1}(v_j)u'(v_j-b(v_j))\frac{db}{dv_j}=0$$

Rearranging yields the following differential equation for $b(v)$

$$\text{(A2)}\quad b'(v)=(n-1)\frac{F'(v)}{F(v)}\frac{u(v-b)}{u'(v-b)}$$

With reserve price $b_0=v_*$ we also have the boundary condition

$$\text{(A3)}\qquad b(v_*)=v_*$$

We wish to compare the solution for two different utility functions, $u_1(\cdot)$ and $u_2(\cdot)$ where the latter exhibits a higher degree of risk aversion, that is,

$$\text{(A4)}\quad -u_2''(x)/u_2'(x)>-u_1''(x)/u_1'(x)\geqslant 0$$

By inspection of (A2), if we can establish that

$$\text{(A5)}\quad \phi(x)=u_2(x)/u_2'(x) -u_1(x)/u_1'(x)>0 \qquad \text{for } x>0$$

then $b_2'(v)>b_1'(v)$ and hence $b_2(v)>b_1(v)$ for all $v>v_*$. To demonstrate (A5) we note first that, since $u(0)=0$ and $u(x)$ is strictly increasing,

$$\text{(A6)}\qquad \frac{u(x)}{u'(x)}>\frac{u(0)}{u'(0)}=0 \qquad \text{for all } x>0$$

Inequality (A5) holds if we can establish that for all x such that $\phi(x)=0, \phi(x)$ is strictly increasing. Differentiating (A5) we have

$$\text{(A7)}\quad \phi'(x)=\left(\frac{-u_2''}{u_2'}\right)\left(\frac{u_2}{u_2'}\right)-\left(\frac{-u_1''}{u_1'}\right)\left(\frac{u_1}{u_1'}\right)$$

From (A4)–(A6), $x>0$ and $\phi(x)=0$ implies that $\phi'(x)>0$. Moreover, differentiating (A7) and setting $x=0$ we also have

$$\phi''(0)>\phi'(0)=0$$

Thus $\phi(x)$ is strictly increasing at $x=0$.

PROPOSITION 5: *Suppose assumption IID holds and all buyers share a common von Neumann-Morgenstern utility function. Then in the high bid auction, the optimal seller reserve price is a declining function of the degree of risk aversion.*

PROOF:

The method of proof is to compare the effect of a change in the reserve price v_* on the equilibrium bid function $b=b(v, v_*)$ for different degrees of risk aversion. Expected seller revenue, $R(v_*)$, is the expected value of the highest ranked bid, that is,

$$R(v_*)=\int_{v_*}^{\bar{v}}b(v,v_*)dF^{n-1}(v)$$

Then the net advantage to the seller if utility is $u_2(\cdot)$ rather than $u_1(\cdot)$ can be expressed as

$$R_2(v_*)-R_1(v_*) = \int_{v_*}^{\bar{v}}\left[b_2(v,v_*)-b_1(v,v_*)\right]dF^{n-1}(v)$$

Differentiating with respect to v_* we have

$$\text{(A8)}\quad R_2'(v_*)-R_1'(v_*) =\int_{v_*}^{\bar{v}}\left[\frac{\partial b_2}{\partial v_*}-\frac{\partial b_1}{\partial v_*}\right]dF^{n-1}(v)$$

It suffices to show that the bracketed expression in (A8) is negative, for then $R_2'(v_*)$ is negative when $R_1'(v_*)$ is zero.

From (A2), the equilibrium bid function $b(v,v_*)$ is the solution to

$$\text{(A9)}\quad \frac{\partial}{\partial v}b(v,v_*)=(n-1)\frac{F'(v)}{F(v)}\frac{u(v-b)}{u'(v-b)}$$

with the boundary condition,

$$\text{(A10)}\qquad b(v_*,v_*)=v_*$$

Assuming $u(\cdot)$ is twice differentiable, we can differentiate (A9) with respect to the reserve price v_* and so obtain the following differential equation for $\partial b/\partial v_*$

$$\text{(A11)}\quad \frac{\partial}{\partial v}\left(\frac{\partial b}{\partial v_*}\right)=-(n-1)\frac{F'(v)}{F(v)} \times\left[1+\left(\frac{-u''}{u'}\right)\left(\frac{u}{u'}\right)\right]\left(\frac{\partial b}{\partial v_*}\right)$$

From (A4) and (A5) the bracket in (A11) is larger for the utility function $u_2(x)$ exhibiting greater risk aversion. Then if we can establish that $\partial b_2/\partial v_*=\partial b_1/\partial v_*>0$ at $v=v_*$, it will follow from (A11) that

$$\frac{\partial}{\partial v}\left(\frac{\partial b_2}{\partial v_*}\right)>\frac{\partial}{\partial v}\left(\frac{\partial b_1}{\partial v_*}\right)$$

for $v>v_*$ and hence that $\partial b_2/\partial v_*>\partial b_1/\partial v_*$ for $v>v_*$.

From (A10) we have,

$$\text{(A12)}\quad \frac{\partial b}{\partial v}(v,v_*)|_{v=v_*}+\frac{\partial b}{\partial v_*}(v,v_*)|_{v=v_*}=1$$

Since $b(v_*,v_*)=v_*$ and $u(0)=0$, it follows from (A2) that for any concave utility function and any $v_*>0$, the first term in (A12) is zero. Then the second term in (A12) is equal to unity for both $u_1(x)$ and $u_2(x)$.

REFERENCES

G. R. Butters, "Equilibrium Price Distributions and the Economics of Information," unpublished doctoral dissertation, Univ. Chicago 1975.

R. Engelbrecht-Wiggans, "Auctions and Bidding Models: A Survey," *Management Sci.*, Feb. 1980, *26*, 119–42.

J. C. Harsanyi, "Games with Incomplete Information Played by Bayesian Players," Parts I; II; III, *Management Sci.*, Nov. 1967; Jan. 1968; Mar. 1968, *14*, 159–82; 321–34; 486–502.

M. Harris and A. Raviv, "Allocation Mechanisms and the Design of Auctions," working paper, Grad. School Ind. Admin., Carnegie-Mellon Univ. 1979.

______ and ______, "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *Amer. Econ. Rev.*, June 1981, *71*, 347–65.

C. A. Holt, Jr., "Competitive Bidding for Contracts under Alternative Auction Procedures," *J. Polit. Econ.*, June 1980, *88*, 433–445.

E. S. Maskin and J. G. Riley, (1980a) "Auctioning an Indivisible Object," working paper, Kennedy School Government, Harvard Univ. 1980.

______ and ______, (1980b) "Price Discrimination and Bundling, Monopoly Selling Strategies when Information is Incomplete," mimeo., MIT, 1980.

S. Matthews, "Risk Aversion and the Efficiency of First and Second Price Auctions," mimeo., Univ. Illinois, 1979.

P. R. Milgrom and R. J. Weber, "A Theory of Auctions and Competitive Bidding," working paper, Grad. School Management, Northwestern Univ. 1980.

R. B. Myerson, "Optimal Auction Design," *Math. Operations Res.*, 1981, forthcoming.

M. E. Oren and A. C. Williams, "On Competitive Bidding," *Operations Research*, Nov.-Dec. 1975, *23*, 1072–79.

A. Ortega-Reichert, "Models for Competitive Bidding Under Uncertainty," unpublished doctoral dissertation, Stanford Univ. 1968.

J. G. Riley and W. F. Samuelson, "Optimal Auctions," working paper, Univ. California-Los Angeles 1979.

W. F. Samuelson, "Models of Competitive Bidding," unpublished doctoral dissertation, Harvard Univ. 1978.

R. M. Stark and M. H. Rothkopf, "Competitive Bidding: A comprehensive Bibliography," *Operations Res.*, Mar.-Apr. 1979, *27*, 364–91.

W. Vickrey, "Counterspeculation, Auctions and Competitive Sealed Tenders," *J. Finance*, Mar. 1961, *16*, 8–37.

R. B. Wilson, "Competitive Bidding with Asymmetrical Information," *Management Sci.*, July 1967, *13*, A816–20.

______, "On the Incentive for Information Acquisition in Competitive Bidding with Asymmetrical Information," report, dept. econ., Stanford Univ. 1975.

Credit Rationing in Markets with Imperfect Information

By JOSEPH E. STIGLITZ AND ANDREW WEISS*

Why is credit rationed? Perhaps the most basic tenet of economics is that market equilibrium entails supply equalling demand; that if demand should exceed supply, prices will rise, decreasing demand and/or increasing supply until demand and supply are equated at the new equilibrium price. So if prices do their job, rationing should not exist. However, credit rationing and unemployment do in fact exist. They seem to imply an excess demand for loanable funds or an excess supply of workers.

One method of "explaining" these conditions associates them with short- or long-term disequilibrium. In the short term they are viewed as *temporary disequilibrium* phenomena; that is, the economy has incurred an exogenous shock, and for reasons not fully explained, there is some stickiness in the prices of labor or capital (wages and interest rates) so that there is a transitional period during which rationing of jobs or credit occurs. On the other hand, long-term unemployment (above some "natural rate") or credit rationing is explained by governmental constraints such as usury laws or minimum wage legislation.[1]

The object of this paper is to show that in *equilibrium* a loan market may be characterized by credit rationing. Banks making loans are concerned about the interest rate they receive on the loan, and the riskiness of the loan. However, the interest rate a bank charges may itself affect the riskiness of the pool of loans by either: 1) sorting potential borrowers (the adverse selection effect); or 2) affecting the actions of borrowers (the incentive effect). Both effects derive directly from the residual imperfect information which is present in loan markets after banks have evaluated loan applications. When the price (interest rate) affects the nature of the transaction, it may not also clear the market.

The adverse selection aspect of interest rates is a consequence of different borrowers having different probabilities of repaying their loan. The expected return to the bank obviously depends on the probability of repayment, so the bank would like to be able to identify borrowers who are more likely to repay. It is difficult to identify "good borrowers," and to do so requires the bank to use a variety of *screening devices*. The interest rate which an individual is willing to pay may act as one such screening device: those who are willing to pay high interest rates may, on average, be worse risks; they are willing to borrow at high interest rates because they perceive their probability of repaying the loan to be low. As the interest rate rises, the average "riskiness" of those who borrow increases, possibly lowering the bank's profits.

Similarly, as the interest rate and other terms of the contract change, the behavior of the borrower is likely to change. For instance, raising the interest rate decreases the return on projects which succeed. We will show that higher interest rates induce firms to undertake projects with lower probabilities of success but higher payoffs when successful.

In a world with perfect and costless information, the bank would stipulate precisely all the actions which the borrower could

*Bell Telephone Laboratories, Inc. and Princeton University, and Bell Laboratories, Inc., respectively. We would like to thank Bruce Greenwald, Henry Landau, Rob Porter, and Andy Postlewaite for fruitful comments and suggestions. Financial support from the National Science Foundation is gratefully acknowledged. An earlier version of this paper was presented at the spring 1977 meetings of the Mathematics in the Social Sciences Board in Squam Lake, New Hampshire.

[1] Indeed, even if markets were not competitive one would not expect to find rationing; profit maximization would, for instance, lead a monopolistic bank to raise the interest rate it charges on loans to the point where excess demand for loans was eliminated.

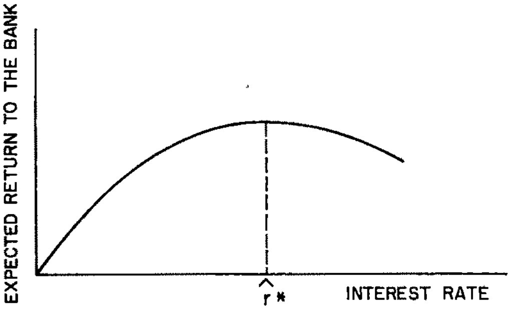


FIGURE 1. THERE EXISTS AN INTEREST RATE WHICH MAXIMIZES THE EXPECTED RETURN TO THE BANK

undertake (which might affect the return to the loan). However, the bank is not able to directly control all the actions of the borrower; therefore, it will formulate the terms of the loan contract in a manner designed to induce the borrower to take actions which are in the interest of the bank, as well as to attract low-risk borrowers.

For both these reasons, the expected return by the bank may increase less rapidly than the interest rate; and, beyond a point, may actually decrease, as depicted in Figure 1. The interest rate at which the expected return to the bank is maximized, we refer to as the "bank-optimal" rate, $\hat{r}^*$.

Both the demand for loans and the supply of funds are functions of the interest rate (the latter being determined by the expected return at $\hat{r}^*$). Clearly, it is conceivable that at $\hat{r}^*$ the demand for funds exceeds the supply of funds. Traditional analysis would argue that, in the presence of an excess demand for loans, unsatisfied borrowers would offer to pay a higher interest rate to the bank, bidding up the interest rate until demand equals supply. But although supply does not equal demand at $\hat{r}^*$, it is the equilibrium interest rate! The bank would not lend to an individual who offered to pay more than $\hat{r}^*$. In the bank's judgment, such a loan is likely to be a worse risk than the average loan at interest rate $\hat{r}^*$, and the expected return to a loan at an interest rate above $\hat{r}^*$ is actually lower than the expected return to the loans the bank is presently making. Hence, there are no competitive forces leading supply to equal demand, and credit is rationed.

But the interest rate is not the only term of the contract which is important. The amount of the loan, and the amount of collateral or equity the bank demands of loan applicants, will also affect both the behavior of borrowers and the distribution of borrowers. In Section III, we show that increasing the collateral requirements of lenders (beyond some point) may decrease the returns to the bank, by either decreasing the average degree of risk aversion of the pool of borrowers; or in a multiperiod model inducing individual investors to undertake riskier projects.

Consequently, it may not be profitable to raise the interest rate or collateral requirements when a bank has an excess demand for credit; instead, banks deny loans to borrowers who are observationally indistinguishable from those who receive loans.[2]

It is not our argument that credit rationing will always characterize capital markets, but rather that it may occur under not implausible assumptions concerning borrower and lender behavior.

This paper thus provides the first theoretical justification of true credit rationing. Previous studies have sought to explain why each individual faces an upward sloping interest rate schedule. The explanations offered are (a) the probability of default for any particular borrower increases as the amount borrowed increases (see Stiglitz 1970, 1972; Marshall Freimer and Myron Gordon; Dwight Jaffee; George Stigler), or (b) the mix of borrowers changes adversely (see Jaffee and Thomas Russell). In these circumstances we would not expect loans of different size to pay the same interest rate, any more than we would expect two borrowers, one of whom has a reputation for prudence and the other a reputation as a bad credit risk, to be able to borrow at the same interest rate.

We reserve the term credit rationing for circumstances in which either (a) among loan applicants who appear to be identical some

[2] After this paper was completed, our attention was drawn to W. Keeton's book. In chapter 3 he develops an incentive argument for credit rationing.

receive a loan and others do not, and the rejected applicants would not receive a loan even if they offered to pay a higher interest rate; or (b) there are identifiable groups of individuals in the population who, with a given supply of credit, are unable to obtain loans at any interest rate, even though with a larger supply of credit, they would.[3]

In our construction of an equilibrium model with credit rationing, we describe a market equilibrium in which there are many banks and many potential borrowers. Both borrowers and banks seek to maximize profits, the former through their choice of a project, the latter through the interest rate they charge borrowers and the collateral they require of borrowers (the interest rate received by depositors is determined by the zero-profit condition). Obviously, we are not discussing a "price-taking" equilibrium. Our equilibrium notion is competitive in that banks compete; one means by which they compete is by their choice of a price (interest rate) which maximizes their profits. The reader should notice that in the model presented below there are interest rates at which the demand for loanable funds equals the supply of loanable funds. However, these are not, in general, equilibrium interest rates. If, at those interest rates, banks could increase their profits by lowering the interest rate charged borrowers, they would do so.

Although these results are presented in the context of credit markets, we show in Section V that they are applicable to a wide class of principal-agent problems (including those describing the landlord-tenant or employer-employee relationship).

I. Interest Rate as a Screening Device

In this section we focus on the role of interest rates as screening devices for distinguishing between good and bad risks. We assume that the bank has identified a group of projects; for each project θ there is a probability distribution of (gross) returns R. We assume for the moment that this distribution cannot be altered by the borrower.

Different firms have different probability distributions of returns. We initially assume that the bank is able to distinguish projects with different mean returns, so we will at first confine ourselves to the decision problem of a bank facing projects having the same mean return. However, the bank cannot ascertain the riskiness of a project. For simplicity, we write the distribution of returns[4] as $F(R,\theta)$ and the density function as $f(R,\theta)$, and we assume that greater θ corresponds to greater risk in the sense of mean preserving spreads[5] (see Rothschild-Stiglitz), i.e., for $\theta_1 > \theta_2$, if

$$(1) \qquad \int_0^\infty R f(R,\theta_1)\,dR = \int_0^\infty R f(R,\theta_2)\,dR$$

then for $y \geqslant 0$,

$$(2) \qquad \int_0^y F(R,\theta_1)\,dR \geqslant \int_0^y F(R,\theta_2)\,dR$$

If the individual borrows the amount B, and the interst rate is $\hat{r}$, then we say the individual defaults on his loan if the return R plus the collateral C is insufficient to pay back the promised amount,[6] i.e., if

$$(3) \qquad C + R \leqslant B(1+\hat{r})$$

[3] There is another form of rationing which is the subject of our 1980 paper: banks make the provision of credit in later periods contingent on performance in earlier period; banks may then refuse to lend even when these later period projects stochastically dominate earlier projects which are financed.

[4] These are subjective probability distributions; the perceptions on the part of the bank may differ from those of the firm.

[5] Michael Rothschild and Stiglitz show that conditions (1) and (2) imply that project 2 has a greater variance than project 1, although the converse is not true. That is, the mean preserving spread criterion for measuring risk is stronger than the increasing variance criterion. They also show that (1) and (2) can be interpreted equally well as: given two projects with equal means, every risk averter prefers project 1 to project 2.

[6] This is not the only possible definition. A firm might be said to be in default if $R < B(1+\hat{r})$. Nothing critical depends on the precise definition. We assume, however, that if the firm defaults, the bank has first claim on $R+C$. The analysis may easily be generalized to include bankruptcy costs. However, to simplify the analysis, we usually shall ignore these costs. Throughout this section we assume that the project is the sole project

Thus the net return to the borrower $\pi(R,\hat{r})$ can be written as

$$\text{(4a)}\quad \pi(R,\hat{r})=max(R-(1+\hat{r})B;-C)$$

The return to the bank can be written as

$$\text{(4b)}\quad \rho(R,\hat{r})=min(R+C;B(1+\hat{r}))$$

that is, the borrower must pay back either the promised amount or the maximum he can pay back $(R+C)$.

For simplicity, we shall assume that the borrower has a given amount of equity (which he cannot increase), that borrowers and lenders are risk neutral, that the supply of loanable funds available to a bank is unaffected by the interest rate it charges borrowers, that the cost of the project is fixed, and unless the individual can borrow the difference between his equity and the cost of the project, the project will not be undertaken, that is, projects are not divisible. For notational simplicity, we assume the amount borrowed for each project is identical, so that the distribution functions describing the number of loan applications are identical to those describing the monetary value of loan applications. (In a more general model, we would make the amount borrowed by each individual a function of the terms of the contract; the quality mix could change not only as a result of a change in the mix of applicants, but also because of a change in the relative size of applications of different groups.)

We shall now prove that the interest rate acts as a screening device; more precisely we establish

THEOREM 1: *For a given interest rate $\hat{r}$, there is a critical value $\hat{\theta}$ such that a firm borrows from the bank if and only if $\theta>\hat{\theta}$.*

This follows immediately upon observing that profits are a convex function of R, as in Figure 2a. Hence expected profits increase with risk.

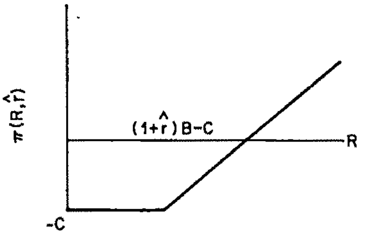


FIGURE 2a. FIRM PROFITS ARE A CONVEX FUNCTION OF THE RETURN ON THE PROJECT

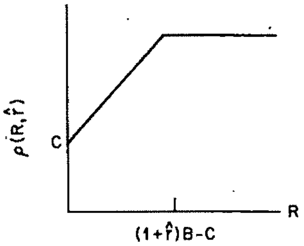


FIGURE 2b. THE RETURN TO THE BANK IS A CONCAVE FUNCTION OF THE RETURN ON THE PROJECT

The value of $\hat{\theta}$ for which expected profits are zero satisfies

$$\text{(5)}\quad \Pi(\hat{r},\hat{\theta})\equiv \int_0^\infty max[R-(\hat{r}+1)B;-C]\,dF(R,\hat{\theta})=0$$

Our argument that the adverse selection of interest rates could cause the returns to the bank to decrease with increasing interest rates hinged on the conjecture that as the interest rate increased, the mix of applicants became worse; or

THEOREM 2: *As the interest rate increases, the critical value of θ, below which individuals do not apply for loans, increases.*

This follows immediately upon differentiating (5):

$$\text{(6)}\quad \frac{d\hat{\theta}}{d\hat{r}}=\frac{B\int_{(1+\hat{r})B-C}^{\infty}dF(R,\hat{\theta})}{\partial\Pi/\partial\hat{\theta}}>0$$

For each θ, expected profits are decreased;

undertaken by the firm (individual) and that there is limited liability. The equilibrium extent of liability is derived in Section III.

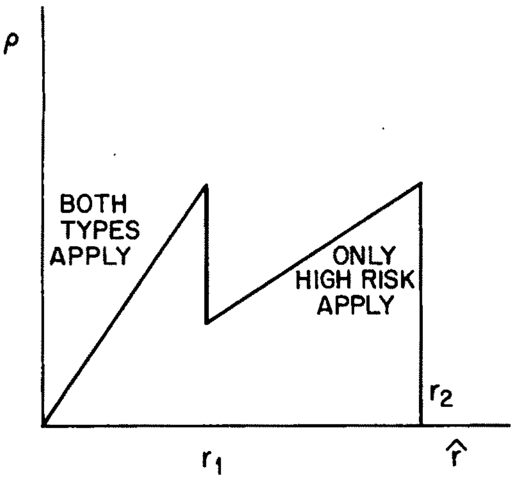


FIGURE 3. OPTIMAL INTEREST RATE r_1

hence using Theorem 1, the result is immediate.

We next show:

THEOREM 3: *The expected return on a loan to a bank is a decreasing function of the riskiness of the loan.*

PROOF:

From (4b) we see that $\rho(R,\hat{r})$ is a concave function of R, hence the result is immediate. The concavity of $\rho(R,\hat{r})$ is illustrated in Figure 2b.

Theorems 2 and 3 imply that, in addition to the usual direct effect of increases in the interest rate increasing a bank's return, there is an indirect, adverse-selection effect acting in the opposite direction. We now show that this adverse-selection effect *may* outweigh the direct effect.

To see this most simply, assume there are two groups; the "safe" group will borrow only at interest rates below r_1, the "risky" group below r_2, and $r_1 < r_2$. When the interest rate is raised slightly above r_1, the mix of applicants changes dramatically: all low risk applicants withdraw. (See Figure 3.) By the same argument we can establish

THEOREM 4: *If there are a discrete number of potential borrowers (or types of borrowers) each with a different θ, $\bar{\rho}(\hat{r})$ will not be a monotonic function of $\hat{r}$, since as each successive group drops out of the market, there is a discrete fall in $\bar{\rho}$ (where $\bar{\rho}(\hat{r})$ is the mean return to the bank from the set of applicants at the interest rate $\hat{r}$).*

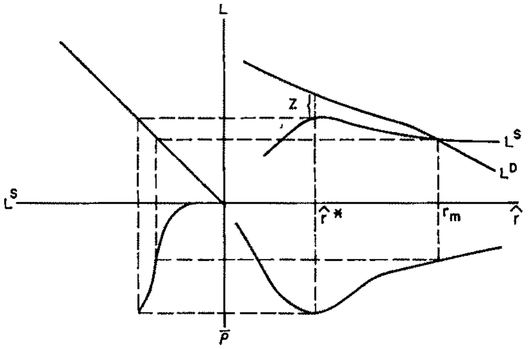


FIGURE 4. DETERMINATION OF THE MARKET EQUILIBRIUM

Other conditions for nonmonotonicity of $\bar{\rho}(\hat{r})$ will be established later. Theorems 5 and 6 show why nonmonotonicity is so important:

THEOREM 5: *Whenever $\bar{\rho}(\hat{r})$ has an interior mode, there exist supply functions of funds such that competitive equilibrium entails credit rationing.*

This will be the case whenever the "Walrasian equilibrium" interest rate—the one at which demand for funds equals supply—is such that there exists a lower interest rate for which $\bar{\rho}$, the return to the bank, is higher.

In Figure 4 we illustrate a credit rationing equilibrium. Because demand for funds depends on $\hat{r}$, the interest rate charged by banks, while the supply of funds depends on ρ, the mean return on loans, we cannot use a conventional demand/supply curve diagram. The demand for loans is a decreasing function of the interest rate charged borrowers; this relation L^D is drawn in the upper right quadrant. The nonmonotonic relation between the interest charged borrowers, and the expected return to the bank per dollar loaned $\bar{\rho}$ is drawn in the lower right quadrant. In the lower left quadrant we depict the relation between $\bar{\rho}$ and the supply of loanable funds L^S. (We have drawn L^S as if it

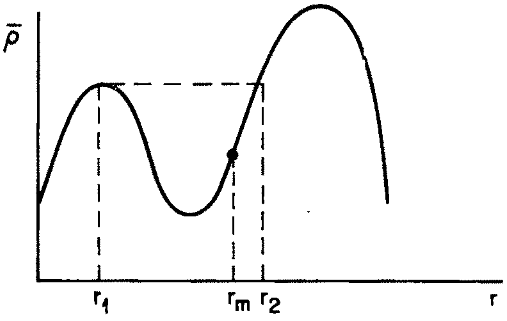


FIGURE 5. A TWO-INTEREST RATE EQUILIBRIUM

were an increasing function of $\bar{p}$. This is not necessary for our analysis.) If banks are free to compete for depositors, then $\bar{p}$ will be the interest rate received by depositors. In the upper right quadrant we plot L^S as a function of $\hat{r}$, through the impact of $\hat{r}$ on the return on each loan, and hence on the interest rate $\bar{p}$ banks can offer to attract loanable funds.

A credit rationing equilibrium exists given the relations drawn in Figure 4; the demand for loanable funds at $\hat{r}^*$ exceeds the supply of loanable funds at $\hat{r}^*$ and any individual bank increasing its interest rate beyond $\hat{r}^*$ would lower its return per dollar loaned. The excess demand for funds is measured by Z. Notice that there is an interest rate r_m at which the demand for loanable funds equals the supply of loanable funds; however, r_m is not an equilibrium interest rate. A bank could increase its profits by charging $\hat{r}^*$ rather than r_m: at the lower interest rate it would attract at least all the borrowers it attracted at r_m and would make larger profits from each loan (or dollar loaned).

Figure 4 can also be used to illustrate an important comparative statics property of our market equilibrium:

COROLLARY 1. *As the supply of funds increases, the excess demand for funds decreases, but the interest rate charged remains unchanged, so long as there is any credit rationing.*

Eventually, of course, Z will be reduced to zero; further increases in the supply of funds then reduce the market rate of interest.

Figure 5 illustrates a $\bar{p}(\hat{r})$ function with multiple modes. The nature of the equilibrium for such cases is described by Theorem 6.

THEOREM 6: *If the $\bar{p}(r)$ function has several modes, market equilibrium could either be characterized by a single interest rate at or below the market-clearing level, or by two interest rates, with an excess demand for credit at the lower one.*

PROOF:

Denote the lowest Walrasian equilibrium interest rate by r_m and denote by $\hat{r}$ the interest rate which maximizes $\rho(r)$. If $\hat{r}<r_m$, the analysis for Theorem 5 is unaffected by the multiplicity of modes. There will be credit rationing at interest rate $\hat{r}$. The rationed borrowers will not be able to obtain credit by offering to pay a higher interest rate.

On the other hand, if $\hat{r}>r_m$, then loans may be made at two interest rates, denoted by r_1 and r_2. r_1 is the interest rate which maximizes $\rho(r)$ conditional on $r\leqslant r_m$; r_2 is the lowest interest rate greater than r_m such that $\rho(r_2)=\rho(r_1)$. From the definition of r_m, and the downward slope of the loan demand function, there will be an excess demand for loanable funds at r_1 (unless $r_1=r_m$, in which case there is no credit rationing). Some rejected borrowers (with reservation interest rates greater than or equal to r_2) will apply for loans at the higher interest rate. Since there would be an excess supply of loanable funds at r_2 if no loans were made at r_1, and an aggregate excess demand for funds if no loans were made at r_2, there exists a distribution of loanable funds available to borrowers at r_1 and r_2 such that all applicants who are rejected at interest rate r_1 and who apply for loans at r_2 will get credit at the higher interest rate. Similarly, all the funds available at $\rho(r_1)$ will be loaned at either r_1 or r_2. (There is, of course, an excess demand for loanable funds at r_1 since every borrower who eventually borrows at r_2 will have first applied for credit at r_1.) There is clearly no incentive for small deviations from r_1, which is a local maximum of $\rho(r)$. A bank lending at an interest rate r_3 such that $\rho(r_3)<\rho(r_1)$ would not be able to obtain credit. Thus, no bank would switch to a loan offer between r_1 and r_2. A bank offering an interest rate r_4 such that $\rho(r_4)>\rho(r_1)$ would not be able to attract any borrowers since by definition $r_4>r_2$, and there is no excess demand at interest rate r_2.

A. *Alternative Sufficient Conditions for Credit Rationing*

Theorem 4 provided a sufficient condition for adverse selection to lead to a nonmonotonic $\bar{\rho}(\hat{r})$ function. In the remainder of this section, we investigate other circumstances under which for some levels of supply of funds there will be credit rationing.

1. *Continuum of Projects*

Let $G(\theta)$ be the distribution of projects by riskiness θ, and $\rho(\theta, r)$ be the expected return to the bank of a loan of risk θ and interest rate r. The mean return to the bank which lends at the interest rate $\hat{r}$ is simply

$$(7)\qquad \bar{\rho}(\hat{r})=\frac{\int_{\hat{\theta}(\hat{r})}^{\infty}\rho(\theta,\hat{r})\,dG(\theta)}{1-G(\hat{\theta})}$$

From Theorem 5 we know that $d\bar{\rho}(\hat{r})/d\hat{r}<0$ for some value of $\hat{r}$ is a sufficient condition for credit rationing. Let $\rho(\hat{\theta},\hat{r})=\hat{\rho}$ so that

$$(8)\quad \frac{d\bar{\rho}}{d\hat{r}}=-\frac{g(\hat{\theta})}{\left[1-G(\hat{\theta})\right]}(\hat{\rho}-\bar{\rho})\frac{d\hat{\theta}}{d\hat{r}} + \frac{\int_{\hat{\theta}}^{\infty}\left[1-F((1+\hat{r})B-C,\theta)\right]dG(\theta)}{1-G(\hat{\theta})}$$

From Theorems 1 and 3, the first term is negative (representing the change in the mix of applicants), while the second term (the increase in returns, holding the applicant pool fixed, from raising the interest charges) is positive. The first term is large, in absolute value, if there is a large difference between the mean return on loans made at interest rate $\hat{r}$ and the return to the bank from the project making zero returns to the firm at interest rate $\hat{r}$ (its "safest" loan). It is also large if $(g(\hat{\theta})/[1-G(\hat{\theta})])$ $(d\hat{\theta}/d\hat{r})$ is large, that is, a small change in the nominal interest rate induces a large change in the applicant pool.

2. *Two Outcome Projects*

Here we consider the simplest kinds of projects (from an analytical point of view), those which either succeed and yield a return R, or fail and yield a return D. We normalize to let $B=1$. All the projects have the same unsuccessful value (which could be the value of the plant and equipment) while R ranges between S and K (where $K>S$). We also assume that projects have been screened so that all projects within a loan category have the same expected yield, T, and there is no collateral required, that is, $C=0$, and if $p(R)$ represents the probability that a project with a successful return of R succeeds, then

$$(9)\qquad p(R)R+\left[1-p(R)\right]D=T$$

In addition, the bank suffers a cost of X per dollar loaned upon loans that default, which could be interpreted as the difference between the value of plant and equipment to the firm and the value of the plant and equipment to the bank. Again the density of project values is denoted by $g(R)$, the distribution function by $G(R)$.

Therefore, the expected return per dollar lent at an interest rate $\hat{r}$, if we let $J=\hat{r}+1$, is (since individuals will borrow if and only if $R>J$):

$$(10)\qquad \rho(J)=\frac{1}{\int_J^K g(R)\,dR}\left[J\int_J^K p(R)g(R)\,dR + \int_J^K[1-p(R)][D-X]g(R)\,dR\right]$$

Using l'Hopital's rule and (1), we can establish sufficient conditions for $lim_{J\rightarrow K}(\partial\rho(J)/\partial J)<0$ (and hence for the nonmonotonicity of ρ):[7]

[7] The proofs of these propositions are slightly complicated. Consider 1. Since $p(R)=T-D/R-D$, the

(a) if $lim_{R\to K}\, g(R)\neq 0,\infty$ then a sufficient condition is $X>K-D$, or equivalently, $lim_{R\to K}\, p(R)+p'(R)X<0$

(b) if $g(K)=0$, $g'(K)\neq 0,\infty$ then a sufficient condition is $2X>K-D$, or equivalently, $lim_{R\to K}\, p(R)+2p'(R)X<0$

(c) if $g(K)=0$, $g'(K)=0$, $g''(K)\neq 0$, then a sufficient condition is $3X>K-K-D$, or equivalently, $lim_{R\to K}\, p(R)+3p'(R)X<0$

Condition (a) implies that if, as $1+\hat{r}\to K$, the probability of an increase in the interest rate being repaid is outweighed by the deadweight loss of riskier loans, the bank will maximize its return per dollar loaned at an interest rate below the maximum rate at which it can loan funds $(K-1)$. The conditions for an interior bank optimal interest rate are significantly less stringent when $g(K)=0$.

3. *Differences in Attitudes Towards Risk*

Some loan applicants are clearly more risk averse than others. These differences will be reflected in project choices, and thus affect the bank-optimal interest rate. High interest rates may make projects with low mean returns— the projects undertaken by risk averse individuals—infeasible, but leave relatively unaffected the risky projects. The mean return to the bank, however, is lower on the riskier projects than on the safe projects. In the following example, it is systematic differences in risk aversion which results in there being an optimal interest rate.

Assume a fraction λ of the population is infinitely risk averse; each such individual undertakes the best perfectly safe project which is available to him. Within that group, the distribution of returns is $G(R)$ where $G(K)=1$. The other group is risk neutral. For simplicity we shall assume that they all face the same risky project with probability of success p and a return, if successful, of $R^*>K$; if not their return is zero. Letting $\hat{R}=(1+\hat{r})B$ the (expected) return to the bank is

(11)
$$\bar{\rho}(\hat{r})=\frac{\left\{\lambda\left(1-G(\hat{R})\right)+(1-\lambda)p\right\}}{\lambda\left(1-G(\hat{R})\right)+(1-\lambda)}(1+\hat{r})$$

$$=\left[1-\frac{(1-p)(1-\lambda)}{\lambda\left(1-G(\hat{R})\right)+(1-\lambda)}\right]\frac{\hat{R}}{B}$$

Hence for $R<K$, the upper bound on returns from the safe project

(12)
$$\frac{d\,ln\,\bar{\rho}}{d\,ln(1+\hat{r})}=1-$$
$$\frac{(1-\lambda)(1-p)\lambda g(\hat{R})\hat{R}}{\left(1-\lambda G(\hat{R})\right)\left(\lambda\left(1-G(\hat{R})\right)+p(1-\lambda)\right)}$$

A sufficient condition for the existence of an interior bank optimal interest rate is again that $lim_{R\to K}\, \partial\bar{\rho}/\partial\hat{r}<0$, or from (12), $\lambda/1-\lambda$ $lim_{R\to K}\, g(R)\hat{R}>p/1-p$. The greater is the riskiness of the risky project (the lower is p), the more likely is an interior bank optimal interest rate. Similarly, the higher is the relative proportion of the risk averse individuals affected by increases in the interest rate to risk neutral borrowers, the more important is

expected profit per dollar loaned may be rewritten as

$$\rho(J)=[J-D+X][T-D]\frac{\int_J^K\frac{g(R)}{R-D}dR}{\int_J^K g(R)\,dR}+D-X$$

Differentiating, and collecting terms

$$\frac{1}{T-D}\frac{\partial\rho}{\partial J}=\frac{\int_J^K\frac{g(R)}{R-D}dR}{\int_J^K g(R)\,dR}+[J-D+X]$$
$$\times\left[\frac{\frac{-g(J)}{J-D}\int_J^K g(R)\,dR+g(J)\int_J^K\frac{g(R)}{R-D}dR}{\left[\int_J^K g(R)\,dR\right]^2}\right]$$

Using l'Hopital's rule and the assumption that $g(K)\neq 0,\infty$

$$\lim_{J\to K}\left(\frac{1}{T-D}\frac{\partial\rho}{\partial J}\right)=\left(\frac{1}{K-D}-\frac{K-D+X}{2(K-D)^2}\right);$$

or $$sign\left(\lim_{J\to K}\frac{1}{T-D}\frac{\partial\rho}{\partial J}\right)=sign(K-D-X)$$

Conditions 2 and 3 follow in a similar manner.

the self-selection effect, and the more likely is an interior bank optimal interest rate.

II. Interest Rate as an Incentive Mechanism

A. *Sufficient Conditions*

The second way in which the interest rate affects the bank's expected return from a loan is by changing the behavior of the borrower. The interests of the lender and the borrower do not coincide. The borrower is only concerned with returns on the investment when the firm does not go bankrupt; the lender is concerned with the actions of the firm only to the extent that they affect the probability of bankruptcy, and the returns in those states of nature in which the firm *does* go bankrupt. Because of this, and because the behavior of a borrower cannot be perfectly and costlessly monitored by the lender, banks will take into account the effect of the interest rate on the behavior of borrowers.

In this section, we show that increasing the rate of interest increases the relative attractiveness of riskier projects, for which the return to the bank may be lower. Hence, raising the rate of interest may lead borrowers to take actions which are contrary to the interests of the lender, providing another incentive for banks to ration credit rather than raise the interest rate when there is an excess demand for loanable funds.

We return to the general model presented above, but now we assume that each firm has a choice of projects. Consider any two projects, denoted by superscripts j and k. We first establish:

THEOREM 7: *If, at a given nominal interest rate r, a risk-neutral firm is indifferent between two projects, an increase in the interest rate results in the firm preferring the project with the higher probability of bankruptcy.*

PROOF:

The expected return to the ith project is given by

$$\pi^i = E\left[max\left(R^i-(1+\hat{r})B, -C\right)\right] \tag{13}$$

so

$$\frac{d\pi^i}{d\hat{r}} = -B(1-F_i((1+\hat{r})B-C)) \tag{14}$$

Thus, if at some $\hat{r}$, $\pi^j=\pi^k$, the increase in $\hat{r}$ lowers the expected return to the borrower from the project with the higher probability of paying back the loan by more than it lowers the expected return from the project with the lower probability of the loan being repaid.

On the other hand, if the firm is indifferent between two projects with the same mean, we know from Theorem 2 that the bank prefers to lend to the safer project. Hence raising the interest rate above $\hat{r}$ could so increase the riskiness of loans as to lower the expected return to the bank.

THEOREM 8: *The expected return to the bank is lowered by an increase in the interest rate at $\hat{r}$ if, at $\hat{r}$, the firm is indifferent between two projects j and k with distributions $F_j(R)$ and $F_k(R)$, j having a higher probability of bankruptcy than k, and there exists a distribution $F_l(R)$ such that*

(a) *$F_j(R)$ represents a mean preserving spread of the distribution $F_l(R)$, and*

(b) *$F_k(R)$ satisfies a first-order dominance relation with $F_l(R)$; i.e., $F_l(R) > F_k(R)$ for all R.*

PROOF:

Since j has a higher probability of bankruptcy than does k, from Theorem 7 and the initial indifference of borrowers between j and k, an increase in the interest rate $\hat{r}$ leads firms to prefer project j to k. Because of (a) and Theorem 3, the return to the bank on a project whose return is distributed as $F_l(R)$ is higher than on project j, and because of (b) the return to the bank on project k is higher than the return on a project distributed as $F_l(R)$.

B. *An Example*

To illustrate the implications of Theorem 8, assume all firms are identical, and have a choice of two projects, yielding, if successful, returns R^a and R^b, respectively (and nothing

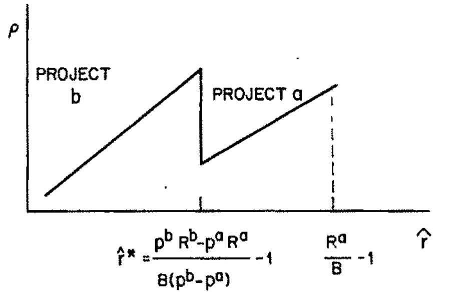


FIGURE 6. AT INTEREST RATES ABOVE $\hat{r}^*$, THE RISKY PROJECT IS UNDERTAKEN AND THE RETURN TO THE BANK IS LOWERED

otherwise) where $R^a > R^b$, and with probabilities of success of p^a and p^b, $p^a < p^b$. For simplicity assume that $C=0$. If the firm is indifferent between the projects at interest rate $\hat{r}$, then

$$(15)\quad \left[R^a - (1+\hat{r})B\right]p^a = \left[R^b - (1+\hat{r})B\right]p^b$$

i.e.,

$$(16)\quad B(1+\hat{r}) = \frac{p^b R^b - p^a R^a}{p^b - p^a} \equiv (1+\hat{r}^*)B$$

Thus, the expected return to the bank as a function of r appears as in Figure 6.

For interest rates below $\hat{r}^*$, firms choose the safe project, while for interest rates between $\hat{r}^*$ and $(R^a/B)-1$, firms choose the risky project. The maximum interest rate the bank could charge and still induce investments in project b is $\hat{r}^*$. The highest interest rate which attracts borrowers is $(R^a/B)-1$, which would induce investment only in project a. Therefore the maximum expected return to a bank occurs when the bank charges an interest rate $\hat{r}^*$ if and only if

$$p^a R^a < \frac{p^b\left(p^b R^b - p^a R^a\right)}{p^b - p^a}.$$

Whenever $p^b R^b > p^a R^a$, $1+\hat{r}^* > 0$, and ρ is not monotonic in $\hat{r}$, so there may be credit rationing.

III. The Theory of Collateral and Limited Liability

An obvious objection to the analysis presented thus far is: When there is an excess demand for funds, would not the bank increase its collateral requirements (increasing the liability of the borrower in the event that the project fails); reducing the demand for funds, reducing the risk of default (or losses to the bank in the event of default) and increasing the return to the bank?

This objection will not in general hold. In this section we will discuss various reasons why banks will not decrease the debt-equity ratio of borrowers (increasing collateral requirements)[8] as a means of allocating credit.

A clear case in which reductions in the debt-equity ratio of borrowers are not optimal for the bank is when smaller projects have a higher probability of "failure," and all potential borrowers have the same amount of equity. In those circumstances, increasing the collateral requirements (or the required proportion of equity finance) of loans will imply financing smaller projects. If projects either succeed or fail, and yield a zero return when they fail, then the increase in the collateral requirement of loans will increase the riskiness of those loans.

Another obvious case where increasing collateral requirements may increase the riskiness of loans is if potential borrowers have different equity, and all projects require the same investment. Wealthy borrowers may be those who, in the past, have succeeded at risky endeavors. In that case they are likely to be less risk averse than the more conservative individuals who have in the past invested in relatively safe securities, and are consequently less able to furnish large amounts of collateral.

In both these examples collateral requirements have adverse selection effects. However, we will present a stronger result. We

[8] Increasing the fraction of the project financed by equity and increasing the collateral requirements both increase the expected return to the bank from any particular project. They have similar but identical risk and incentive effects. Although the analysis below focuses on collateral requirements, similar arguments apply to dept-equity ratios.

will show that even if there are no increasing returns to scale in production and all individuals have the same utility function, the sorting effect of collateral requirements can still lead to an interior bank-optimal level of collateral requirements similar to the interior bank-optimal interest rate derived in Sections I and II. In particular, since wealthier individuals are likely to be less risk averse, we would expect that those who could put up the most capital would also be willing to take the greatest risk. We show that this latter effect is sufficiently strong that increasing collateral requirements will, under plausible conditions, lower the bank's return.

To see this most clearly, we assume all borrowers are risk averse with the same utility function $U(W)$, $U'>0$, $U''<0$. Individuals differ, however, with respect to their initial wealth, W_0. Each "entrepreneur" has a set of projects which he can undertake; each project has a probability of success $p(R)$, where R is the return if successful. If the project is unsuccessful, the return is zero; $p'(R)<0$. Each individual has an alternative safe investment opportunity yielding the return ρ^*. The bank cannot observe either the individual's wealth or the project undertaken. It offers the same contract, defined by C, the amount of collateral, and $\hat{r}$, the interest rate, to all customers. The analysis proceeds as earlier; we first establish:

THEOREM 9: *The contract* $\{C, \hat{r}\}$ *acts as a screening mechanism: there exist two critical values of* W_0, $\hat{W}_0$, *and* $\hat{\hat{W}}_0$, *such that if there is decreasing absolute risk aversion all individuals with wealth* $\hat{W}_0 < W_0 < \hat{\hat{W}}_0$ *apply for loans.*

PROOF:

As before, we normalize so that all projects cost a dollar. If the individual does not borrow, he either does not undertake the project, obtaining a utility of $U(W_0\rho^*)$, or he finances it all himself, obtaining an expected utility of (assuming $W_0 \geqslant 1$)

$$(17)\quad \max_R \{U((W_0-1)\rho^*+R)p(R) + U((W_0-1)\rho^*)(1-p(R))\} \equiv \hat{V}(W_0)$$

Define

$$(18)\quad V_0(W_0)=max\{U(W_0\rho^*), \hat{V}(W_0)]$$

We note that

$$(19)\quad \frac{dU(W_0\rho^*)}{dW_0}=U'\rho^*$$

$$(20)\quad \frac{d\hat{V}(W_0)}{dW_0}=[U_1'p+U_2'(1-p)]\rho^*$$

(where the subscript 1 refers to the state "success" and the subscript 2 to the state "failure"). We can establish that if there is decreasing absolute risk aversion,[9]

$$\frac{dU(W_0\rho^*)}{dW_0}<\frac{d\hat{V}(W_0)}{dW_0}$$

Hence, there exists a critical value of W_0, $\hat{\hat{W}}_0$, such that if $W_0 > \hat{\hat{W}}_0$ individuals who do not borrow undertake the project.

For the rest of the analysis we confine ourselves to the case of decreasing absolute risk aversion and wealth less than $\hat{\hat{W}}_0$.

If the individual borrows, he attains a utility level[10]

$$(21)\quad \{\max_R U(W_0\rho^*-(1+\hat{r})+R)p + U((W_0-C)\rho^*)(1-p)\} \equiv V_B(W_0)$$

The individual borrows if and only if

$$(22)\quad V_B(W_0) \geqslant V_0(W_0)$$

[9] To prove this, we define $\hat{\hat{W}}_0$ as the wealth where undertaking the risky project is a mean-utility preserving spread (compare Peter Diamond-Stiglitz) of the safe project. But writing $U'(W(U))$, where $W(U)$ is the value of terminal wealth corresponding to utility level U,

$$\frac{dU'}{dU}=\frac{U''}{U'}=-A;\quad \frac{d^2U'}{dU^2}=-\frac{A'}{U'}\gtrless 0 \text{ as } A'\lessgtr 0$$

Hence with decreasing absolute risk aversion, U' is a convex function of U and therefore EU' for the risky investment exceeds $U'(\rho^* W_0)$.

[10] In this formulation, the collateral earns a return ρ^*.

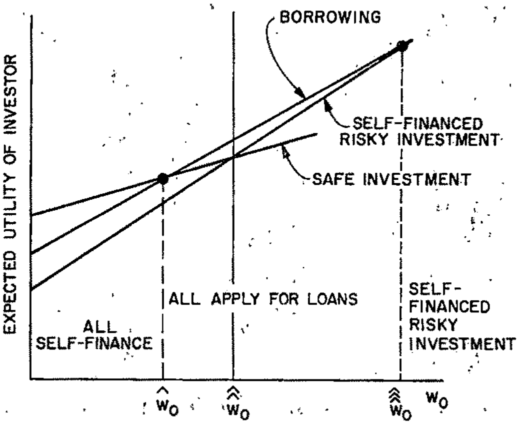


FIGURE 7. COLLATERAL SERVES AS A SCREENING DEVICE

But

$$(23)\qquad \frac{dV_B}{dW_0}=(U_1'p+U_2'(1-p))\rho^*$$

Clearly, only those with $W_0>C$ can borrow. We assume there exists a value of $W_0>0$, denoted $\hat{W}_0$, such that $V_B(\hat{W}_0)=U(\rho^*\hat{W}_0)$. (This will be true for some values of ρ^*.) By the same kind of argument used earlier, it is clear that at $\hat{W}_0$, borrowing with collateral is a mean-utility preserving spread of terminal wealth in comparison to not borrowing and not undertaking the project. Thus using (20) and (23), $dV_B/dW_0>dV_0(W_0)/dW_0$ at $\hat{W}_0$. Hence, for $\hat{W}_0<W_0<\hat{\hat{W}}_0$ all individuals apply for loans, as depicted in Figure 7. Thus, restricting ourselves to $W_0<\hat{\hat{W}}_0$, we have established that if there is any borrowing, it is the wealthiest in that interval who borrow. (The restriction $W_0<\hat{\hat{\hat{W}}}_0$ is weaker than the restriction that the scale of projects exceeds the wealth of any individual.)

Next, we show:

THEOREM 10: *If there is decreasing absolute risk aversion, wealthier individuals undertake riskier projects*: $dR/dW_0>0$.

PROOF:

From (21), we obtain the first-order condition for the choice of R:

$$(24)\qquad U_1'p+(U_1-U_2)p'=0$$

so, using the second-order conditions for a maximum, and (24),

$$(25)\qquad \frac{dR}{dW_0}\gtrless 0 \text{ as } \frac{U_1''p+(U_1'-U_2')p'}{U_1'p}$$

$$=-A_1-\frac{(U_1'-U_2')}{U_1-U_2}\gtrless 0$$

But

$$\lim_{W_1\to W_2}-\frac{U_1'-U_2'}{U_1-U_2}=-\frac{U_1''}{U_1'}=A_1$$

implying that, if $W_1=W_2$, $dR/dW_0=0$. However,

$$\left.\frac{\partial\left(-A_1-\dfrac{U_1'-U_2'}{U_1-U_2}\right)}{\partial W_1}\right|_{A_1=-\frac{U_2'-U_1'}{U_1-U_2}}$$

$$=-A_1'-\frac{U_1''}{U_1-U_2}+\frac{U_1'-U_2'}{U_1-U_2}\frac{U_1'}{U_1-U_2}$$

$$=-A_1'\gtrless 0 \text{ as } A_1'\lesseqgtr 0$$

Hence $dR/dW_0>0$ if $A'<0$.

Next we show

THEOREM 11: *Collateral increases the bank's return from any given borrower*:

$$dp/dC>0$$

PROOF:

This follows directly from the first-order condition (24):

$$sign\,\frac{dR}{dC}=sign\,U_2'\rho^*p'<0$$

and thus $dp/dC>0$. But

THEOREM 12: *There is an adverse selection effect from increasing the collateral requirement, i.e., both the average and the marginal borrower who borrows is riskier,*[11] $d\hat{W}_0/dC>0$:

[11] At a sufficiently high collateral, the wealthy individual will not borrow at all.

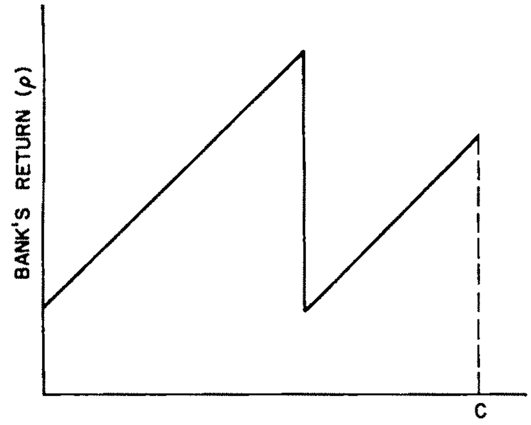


FIGURE 8. INCREASING COLLATERAL REQUIREMENT LOWERS BANK'S RETURNS

PROOF:

This follows immediately upon differentiation of (21)

$$dV_B/dC = -U_2'p^*(1-p) < 0$$

It is easy to show now that this adverse selection effect *may* more than offset the positive direct effect. Assume there are two groups; for low wealth levels, increasing C has no adverse selection effect, so returns are unambiguously increased; but there is a critical level of C such that requiring further investments select against the low wealth-low risk individuals, and the bank's return is lowered.[12] (See Figure 8.)

This simple example has demonstrated[13] that although collateral may have beneficial incentive effects, it may also have countervailing adverse selection effects.

A. *Adverse Incentive Effects*

Although in the model presented above, increasing collateral has a beneficial incentive effect, this is not necessarily the case. The bank has limited control over the actions of the borrowers, as we noted earlier. Thus, the response of the borrower to the increase in lending may be to take actions which, in certain contingencies, will require the bank to lend more in the future. (This argument seems implicit in many discussions of the importance of adequate initial funding for projects.) Consider, for instance, the following simplified multiperiod model. In the first period, θ occurs with probability p_1; if it does, the return to the project (realized the second period) is R_1. If it does not, either an additional amount M must be invested, or the project fails completely (has a zero return). If the bank charges an interest rate $r_2 \leqslant \hat{r}_2$ on these additional funds, they will invest them in "safe" ways; if $r_2 > \hat{r}_2$ those funds will be invested in risky ways. Following the analysis in Section II, we assume that the risk differences are sufficiently strong that the bank charges $\hat{r}_2$ for additional funds. Assume that there is also a set of projects (actions) which the firm can undertake in the first period, but among which the bank cannot discriminate. The individual has an equity of a dollar, which he cannot raise further, so the effect of a decrease in the loan is to affect the actions which the individual takes, that is, it affects the parameters of the projects, R_1, R_2, and M, where M is the amount of second-period financing needed if the project fails in the first period. For simplicity, we take R_2 as given, and let L be the size of the first-period loan. Thus the expected return to the firm is simply (if the additional loan M is made when needed)

$$p_1\left(R_1-(1+\hat{r}_1)^2L\right)$$
$$+\hat{p}\left(R_2-\left[(1+\hat{r}_1)^2L+(1+\hat{r}_2)M\right]\right)$$

where $\hat{p}=p_2(1-p_1)$, $(1+\hat{r}_1)^2$ is the amount paid back (per dollar borrowed) at the end of the second period on the initial loan and $\hat{r}_2$ is the interest on the additional loan M; thus the firm chooses R_1 so that

$$p_1=\hat{p}(1+\hat{r}_2)\frac{dM}{dR_1}$$

[12] If we had not imposed the restriction $W_0 < \hat{\hat{W}}_0$, then there may exist a value of W_0, $\hat{\hat{\hat{W}}}_0 > \hat{\hat{W}}_0$, such that for $W_0 > \hat{\hat{\hat{W}}}_0$, individuals self-finance. It is easy to show that $\partial\hat{\hat{\hat{W}}}_0/\partial C < 0$, so there is a countervailing positive selection effect. However if the density distribution of wealth is decreasing fast enough, then the adverse selection effect outweighs the positive selection effect.

[13] It also shows that the results of earlier sections can be extended to the risk averse entrepreneur.

Assume that the opportunity cost of capital to the bank per period is ρ^*. Then its net expected return to the loan is

$$p_1(1+\hat{r}_1)^2L+\hat{p}\left[(1+\hat{r}_1)^2L+(1+\hat{r}_2)M\right]$$
$$-\rho^*\left[\rho^*L+(1-p_1)M\right]$$

We can show that under certain circumstances, it will pay the bank to extend the line of credit M. Thus, although the bank controls L, it does not control directly the total (expected value) of its loans per customer, $L+(1-p_1)M$.

But more to the point is the fact that the expected return to the bank may not be monotonically decreasing in the size of the first-period loans. For instance, under the hypothesis that $\hat{r}_1$ and $\hat{r}_2$ are optimally chosen and at the optimum $\rho^*>p_2(1+\hat{r}_2)$, the return to the bank is a decreasing function of M/L. Thus, if the optimal response of the firm to a decrease in L is an increase in M (or a decrease in M so long as the percentage decrease in M is less than the percentage decrease in L), a decrease in L actually lowers the bank's profits.[14]

IV. Observationally Distinguishable Borrowers

Thus far we have confined ourselves to situations where all borrowers appear to be identical. Let us now extend the analysis to the case where there are n observationally distinguishable groups each with an interior bank optimal interest rate denoted by r_i^*.[15] The function $\rho_i(r_i)$ denote the gross return to a bank charging a type i borrower interest r_i. We can order the groups so that for $i>j$, $max\, \rho_i(\hat{r}_i)>max\, \rho_j(\hat{r}_j)$.

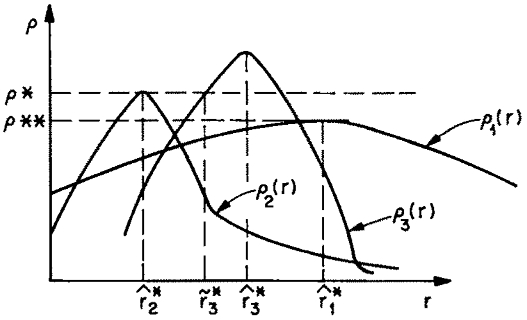


FIGURE 9. IF GROUPS DIFFER, THERE WILL EXIST RED LINING

THEOREM 13: *For $i>j$, type j borrowers will only receive loans if credit is not rationed to type i borrowers.*

PROOF:

Assume not. Since the maximum return on the loan to j is less than that to i, the bank could clearly increase its return by substituting a loan to i for a loan to j; hence the original situation could not have been profit maximizing.

We now show

THEOREM 14: *The equilibrium interest rates are such that for all i, j receiving loans, $\rho_i(\hat{r}_i)=\rho_j(\hat{r}_j)$.*

PROOF:

Again the proof is by contradiction. Let us assume that $\rho_i(\hat{r}_i)>\rho_j(\hat{r}_j)$; then a bank lending to type j borrowers would prefer to bid type i borrowers away from other banks. If ρ^* is the equilibrium return to the banks per dollar loaned, equal to the cost of loanable funds if banks compete freely for borrowers, then for all i, j receiving loans $\rho_i(r_i)=\rho_j(r_j)=\rho^*$. These results are illustrated for three types of borrowers in Figure 9.

If banks have a cost of loanable funds ρ^* then no type 1 borrower will obtain a loan; all type 3 borrowers wishing to borrow at interest rate $\tilde{r}_3$ (which is less than $\hat{r}_3^*$, the rate which maximizes the bank's return) will obtain loans—competition for those borrowers drives their interest rate down; while some, but not necessarily all, type 2 borrowers re-

[14] For instance, if some of the initial investment is for "back-up" systems in case of various kinds of failure, if the reduction in initial funding leads to a reduction in investment in these back-up systems, when a failure does occur, large amounts of additional funding may be required.

[15] The analysis in this section parallels Weiss (1980) in which it was demonstrated that market equilibrium could result in the exclusion of some groups of workers from the labor market.

ceive a loan at $\hat{r}_2^*$. If the interest rate were to fall to ρ^{**}, then all types 2 and 3 would receive loans; and some (but not all) type 1 borrowers would be extended credit.

Groups such as type 1 which are excluded from the credit market may be termed "redlined" since there is no interest rate at which they would get loans if the cost of funds is above ρ^{**}. It is possible that the investments of type 1 borrowers are especially risky so that, although $\rho_1(\hat{r}_1^*)<\rho_3(\hat{r}_3^*)$, the total expected return to type 1 investments (the return to the bank plus the return to the borrower) exceeds the expected return to type 3 investments. It may also be true that type 1 loans are unprofitable to the bank because they find it difficult to filter out risky type 1 investments. In that case it is possible that the return to the bank to an investment by a type 1 borrower would be greater than the return to a type 3 investment if the bank could exercise the same control (judgment) over each group of investors.

Another reason for $\rho_1(\hat{r}_1^*)<\rho_3(\hat{r}_3^*)$ may be that type 1 investors have a broader range of available projects. They can invest in all the projects available to type 3 borrowers, but can also invest in high-risk projects unavailable to type 3. Either because of the convexity of the profit function of borrowers, or because riskier investments have higher expected returns type 1 borrowers will choose to invest in these risky projects.

Thus, *there is no presumption that the market equilibrium allocates credit to those for whom the expected return on their investments is highest.*

IV. Debt vs. Equity Finance, Another View of the Principal-Agent Problem

Although we have phrased this paper in the context of credit markets, the analysis could apply equally well to any one of a number of principal-agent problems. For example, in agriculture the bank (principal) corresponds to the landlord and the borrower (agent) to the tenant while the loan contract corresponds to a rental agreement. The return function for the landlord and tenant appears in Figures 10a and 10b. The central concern in those principal-agent problems is how to provide the proper incentives for the agent. In general, revenue sharing arrangements such as equity finance, or sharecropping are inefficient. Under those schemes the managers of a firm or the tenant will equate their marginal disutility of effort with their share of their marginal product rather than with their total marginal product. Therefore, too little effort will be forthcoming from agents.

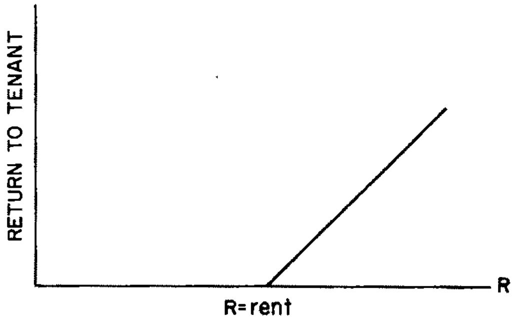


FIGURE 10a

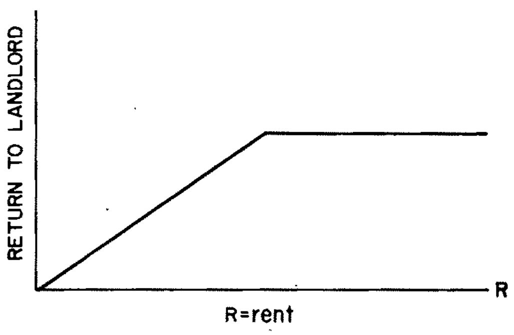


FIGURE 10b

Fixed-fee contracts (for example, rental agreements in agriculture, loan contracts in credit markets) have the disadvantage that they impose a heavy risk on the agent, and thus if agents are risk averse, they may not be desirable. But it has long been thought that they have a significant advantage in not distorting incentives and thus if the agent is risk neutral, fixed-fee contracts will be employed.[16] These discussions have not consid-

[16] See, for instance, Stiglitz (1974). For a recent formalization of the principal-agent problem, see Steven Shavell.

ered the possibility that the agent will fail to pay the fixed fee. In the particular context of the bank-borrower relationship, the assumption that the loan will always be repaid (with interest) seems most peculiar. A borrower can repay the loan in all states of nature only if the risky project's returns plus the value of the equilibrium level of collateral exceeds the safe rate of interest in all states of nature.

The consequences of this are important. Since the agent can by his actions affect the probability of bankruptcy, fixed-fee contracts do not eliminate the incentive problem.

Moreover, they do not necessarily lead to optimal resource allocations. For example, in the two-project case discussed above (Section II, Part B), if expected returns to the safe project exceed that to the risky ($p^sR^s > p^rR^r$) but the highest rate which the bank can charge consistent with the safe project being chosen (r^*) is too low (i.e., $p^s(1+r^*) > p^rR^r$) then the bank chooses an interest rate which causes all its loans to be for risky projects, although the expected total (social) returns on these projects are less than on the safe projects. In this case a usury law forbidding interest rates in excess of r^* will increase net national output. Our 1980 paper and Janusz Ordover and Weiss show that government interventions of various forms lead to Pareto improvements in the allocation of credit.

Because neither equity finance nor debt finance lead to efficient resource allocations, we would not expect to see the exclusive use of either method of financing (even with risk-neutral agents and principals). Similarly, in agriculture, we would not expect to see the exclusive use of rental or sharecropping tenancy arrangements. In general, where feasible, the payoff will be a non-linear function of output (profits). The terms of these contracts will depend on the risk preferences of the principal and agent, the extent to which their actions (both the level of effort and riskiness of outcomes) can affect the probability of bankruptcy, and actions can be specified within the contract or controlled directly by the principal.

One possible criticism of this paper is that the single period analysis presented above artificially limits the strategy space of lenders. In a multiperiod context, for instance, banks could reward "good" borrowers by offering to lend to them at lower interest rates, and this would induce firms to undertake safer projects (just as in the labor market, the promise of promotion and pay increases is an important part of the incentive and sorting structure of firms, see Stiglitz, 1975, J. L. Guasch and Weiss, 1980, 1981). In our 1980 paper, we analyze the nature of equilibrium contracts in a dynamic context. We show that such contingency contracts may characterize the dynamic equilibrium. Indeed, we establish that the bank may want to use quantity constraints—the availability of credit—as an additional incentive device; thus, in the dynamic context there is a further argument for the existence of rationing in a competitive economy.

Even after introducing all of these additional instruments (collateral, equity, nonlinear payment schedules, contingency contracts) there may exist a contract which is optimal from the point of view of the principal; he will not respond, then, to an excess supply of agents by altering the terms of that contract; and there may then be rationing of the form discussed in this paper, that is, an excess demand for loans (capital, land) at the "competitive" contract.

VI. Conclusions

We have presented a model of credit rationing in which among observationally identical borrowers some receive loans and others do not. Potential borrowers who are denied loans would not be able to borrow even if they indicated a willingness to pay more than the market interest rate, or to put up more collateral than is demanded of recipients of loans. Increasing interest rates or increasing collateral requirements could increase the riskiness of the bank's loan portfolio, either by discouraging safer investors, or by inducing borrowers to invest in riskier projects, and therefore could decrease the bank's profits. Hence neither instrument will necessarily be used to equate the supply of loanable funds with the demand for loanable funds. Under those circumstances credit restrictions take the form of limiting the number of loans the bank will make, rather than limiting the size of each loan, or making the interest rate charged an increasing function of the magnitude of the loan, as in most previous discussions of credit rationing.

Note that in a rationing equilibrium, to the extent that monetary policy succeeds in shifting the supply of funds, it will affect the level of investment, not through the interest rate mechanism, but rather through the availability of credit. Although this is a "monetarist" result, it should be apparent that the mechanism is different from that usually put forth in the monetarist literature.

Although we have focused on analyzing the existence of excess demand equilibria in credit markets, imperfect information can lead to excess supply equilibria as well. We will sketch an outline of an argument here (a fuller discussion of the issue and of the macro-economic implications of this paper will appear in future work by the authors in conjunction with Bruce Greenwald).[17] Let us assume that banks make higher expected returns on some of their borrowers than on others: they know who their most credit worthy customers are, but competing banks do not. If a bank tries to attract the customers of its competitors by offering a lower interest rate, it will find that its offer is countered by an equally low interest rate when the customer being competed for is a "good" credit risk, and will not be matched if the borrower is not a profitable customer of the bank. Consequently, banks will seldom seek to steal the customers of their competitors, since they will only succeed in attracting the least profitable of those customers (introducing some noise in the system enables the development of an equilibrium). A bank with an excess supply of loanable funds must assess the profitability of the loans a lower interest rate would attract. In equilibrium each bank may have an excess supply of loanable funds, but no bank will lower its interest rate.

The reason we have been able to model excess demand and excess supply equilibria in credit markets is that the interest rate directly affects the quality of the loan in a manner which matters to the bank. Other models in which prices are set competitively and non-market-clearing equilibria exist share the property that the expected quality of a commodity is a function of its price (see Weiss, 1976, 1980, or Stiglitz, 1976a, b for the labor market and C. Wilson for the used car market).

In any of these models in which, for instance, the wage affects the quality of labor, if there is an excess supply of workers at the wage which minimizes labor costs, there is not necessarily an inducement for firms to lower wages.

The Law of Supply and Demand is not in fact a law, nor should it be viewed as an assumption needed for competitive analysis. It is rather a result generated by the underlying assumptions that prices have neither sorting nor incentive effects. The usual result of economic theorizing: that prices clear markets, is model specific and is not a general property of markets—unemployment and credit rationing are not phantasms.

[17] A similar argument to that presented here appears in Greenwald in the context of labor markets.

REFERENCES

P. Diamond and J. E. Stiglitz, "Increases in Risk and in Risk Aversion," *J. Econ. Theory*, July 1974, *8*, 337–60.

M. Freimer and M. J. Gordon, "Why Bankers Ration Credit," *Quart. J. Econ.*, Aug. 1965, *79*, 397–416.

Bruce Greenwald, *Adverse Selection in the Labor Market*, New York: Garland Press 1979.

J. L. Guasch and A. Weiss, "Wages as Sorting Mechanisms: A Theory of Testing," *Rev. Econ. Studies*, July 1980, *47*, 653–65.

______ **and** ______, "Self-Selection in the Labor Market," *Amer. Econ. Rev.*, forthcoming.

Dwight Jaffee, *Credit Rationing and the Commercial Loan Market*, New York: John Wiley & Sons 1971.

______ **and T. Russell**, "Imperfect Information and Credit Rationing," *Quart. J. Econ.* Nov. 1976, *90*, 651–66.

W. Keeton, *Equilibrium Credit Rationing*, New York: Garland Press 1979.

J. Ordover and A. Weiss, "Information and the Law: Evaluating Legal Restrictions on Competitive Contracts," *Amer. Econ. Rev. Proc.*, May 1981, *71*, 399–404.

M. Rothschild and J. E. Stiglitz, "Increasing Risk: I, A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

S. Shavell, "Risk Sharing and Incentives in the Principal and Agent Problem," *Bell J. Econ.*, Spring 1979, *10*, 55–73.

G. Stigler, "Imperfections in the Capital Market," *J. Polit. Econ.*, June 1967, *85*, 287–92.

J. E. Stiglitz, "Incentives and Risk Sharing in Sharecropping," *Rev. Econ. Studies*, Apr. 1974, *41*, 219–55.

______, "Incentives, Risk, and Information: Notes Towards a Theory of Hierarchy," *Bell J. Econ.*, Autumn 1975, *6*, 552–79.

______, "Prices and Queues as Screening Devices in Competitive Markets," IMSSS tech. report no. 212, Stanford Univ.

______, "The Efficiency Wage Hypothesis, Surplus Labor and the Distribution of Income in L.D.C.'s," *Oxford Econ. Papers*, July 1976, *28*, 185–207.

______, "Perfect and Imperfect Capital Markets," paper presented to the New Orleans meeting of the Econometric Society, Dec. 1970.

______, "Some Aspects of the Pure Theory of Corporate Finance: Bankruptcies and Take-Overs," *Bell J. Econ.*, Autumn 1972, *3*, 458–82.

______ **and A. Weiss,** "Credit Rationing in Markets with Imperfect Information, Part II: A Theory of Contingency Contracts," mimeo. Bell Laboratories and Princeton Univ. 1980.

A. Weiss, "A Theory of Limited Labor Markets," unpublished doctoral dissertation, Stanford Univ. 1976.

______, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *J. Polit. Econ.*, June 1980, *88*, 526–38.

C. Wilson, "The Nature of Equilibrium in Markets with Adverse Selection," *Bell J. Econ.*, Spring 1980, *11*, 108–30.

The Generational Optimum Economy: Extracting Monopoly Gains from Posterity Through Taxation of Capital

By LAWRENCE D. KROHN*

It is well known that an income tax, indeed any tax that falls on capital, often produces a distortion by reducing the effective interest rate below capital's marginal product. This distortion is usually deemed undesirable. This paper demonstrates, however, that such a tax, in an economy constrained to make no intergenerational transfers, can in spite of distortion redound to the benefit of those on whom it is levied. The tax of optimal magnitude is shown to yield a gain which is extracted at the expense of posterity, and which constitutes the exercise of monopoly power by living generations over those yet unborn. Finally it is demonstrated that an economy characterized by such exercise of monopoly power by each generation ad infinitum may produce a utility stream dominating that of an otherwise identical *laissez-faire* economy.

The paper assumes, unrealistically, that generations are aware of this potential gain and that the political process is sufficiently responsive to translate the electorate's desires into taxes of the optimal size. Each generation is assumed represented by its own government during its years of positive capital accumulation. Factor pricing is assumed competitive, that is, according to marginal productivity. Individuals are egoistic and society's distributional ethic enjoins a generation (through its government) from taxing any other generation. There is no government spending on goods and services.

Serving as the engine of this analysis is the discrete-time, infinite-horizon, overlapping-generations apparatus developed by Paul Samuelson and Peter Diamond.[1] Section I establishes the necessary and sufficient conditions for the optimum of each generation, compares this optimum to the *laissez-faire* solution and shows how the optimum can be implemented through an interest income tax. Section II derives the dynamic and steady-state implications of sequential optimization by each generation ad infinitum and compares the economy thus characterized, denoted a generational optimum (GO) economy, to the *laissez-faire* economy with identical initial capital-labor ratio, production and utility functions. Some concluding remarks follow.

Within the framework of the Samuelson-Diamond model, the following are defined for the representative individual of generation t:

c_t = first period consumption
x_t = second period consumption
$u_t(c_t, x_t)$ = the lifetime utility function
w_t = first period (wage) income
s_t = saving $\equiv w_t - c_t$
c_w, s_w = marginal propensities to consume and save out of wages
c_r, c_r^* = the uncompensated and compensated price effects of r on c.

The following are macro-economic variables:

K_t = the aggregate capital stock at period t
L_t = the (fully employed) labor force at period t
k_t = the ratio K/L at period t ($\equiv s_{t-1}/[1+g]$)

*Université Laval. This is an abbreviated version of an earlier paper (1978b) which was based on the third chapter of my doctoral dissertation. While greatly indebted to Edmund Phelps, Guillermo Calvo, George Borts, and an anonymous referee for their constructive suggestions, I alone bear responsibility for errors and shortcomings.

[1] Each generation lives two production periods: the first of work/saving, the second of retirement/dissaving. The two-period life cycle saving model is, of course, that of Irving Fisher.

r = the net (after-tax) interest rate on savings.
g = the constant, exogenous population (and labor force) growth rate
$f(k)$ = the intensive aggregative production function in gross form[2]
$e(k)=f'(k)+kf''(k)$ [the first derivative of $kf'(k)$]
$m(k)=2f''(k)+kf'''(k)$ [the second derivative of $kf'(k)$]

The following assumptions are used throughout:

(a) $u(c,x)\in C^2$ *on the nonnegative orthant*
(b) $[y_c \quad y_x]\begin{bmatrix} u_{cc} & u_{cx} \\ u_{cx} & u_{xx}\end{bmatrix}[y_c \quad y_x]<0$ *for nonzero* y *satisfying* $y_cu_c+y_xu_x=0$[3, 4]
(c) $u_c>0$, $u_x>0$ *on the nonnegative orthant*
(d) $lim_{c\to 0}u_c=\infty$ *all* $x\geqslant 0$, $lim_{x\to 0}u_x=\infty$ *all* $c\geqslant 0$
(e) $0<c_w<1$ *on the positive orthant*
(f) *All consumers over all time have the same preference structure and consumers of any one generation are identically situated in every way at all times*
(g) $f(k)\in C^3$ *over all nonnegative* k
(h) $f(k)>0$ *over all positive* k
(i) $f'(k)>0$ *over all nonnegative* k
(j) $f''(k)<0$ *over all nonnegative* k
(k) $m(k)<0$ *over all nonnegative* k
(l) $k_0>0$, $L_0>0$, $1+g\equiv L_t/L_{t-1}>0$
(m) $kf'(k)=0$ *when* $k=0$

I. The Generational Optimum Economy

A. *The Model*

Suppose that, instead of responding atomistically to the rate of interest facing them, young savers determine the optimal amount of saving they should make from a collective perspective, and suppose too that they can implement their decision through a government that represents them. Their choice is affected in two ways by society's distributional ethic: first, they have available to spend or save no more than the market's reward—the marginal product of their own labor; second, when old they can anticipate receipt of no more or less than the gross marginal product of their capital (savings), that is, the net marginal product of capital plus their original investment.[5]

Mathematically they are then solving the following problem, framed in terms of the representative young person:

$$(1)\qquad \max_{c,x,k_t} u(c,x)$$

$$\text{subject to}\quad c+(1+g)k_t-w(k_{t-1})=0$$
$$x-(1+g)k_tf'(k_t)=0$$

where $w(k_{t-1})\equiv f(k_{t-1})-k_{t-1}f'(k_{t-1})$.

In Appendix B, the solution of this problem is seen to yield a first-order condition of $u_c=e(k_t)u_x$. The expression $e(k_t)$ represents the marginal return per generation t person to an additional period t capital-labor unit, that is, to L_t additional units of capital. The second-order condition is $-Z+u_xm(k_t)/(1+g)<0$. (Z is defined in fn. 4.)

The comparative statics analysis in Appendix B yields the following expression for the behavior of k over time:

$$(2)\quad dk_t/dk_{t-1}$$
$$=-k_{t-1}f''_{t-1}s_w/[1+g+c_r^*m(k_t)]$$

By strict quasi concavity of utility (Assumption (b)), c_r^* is negative and by Assumptions (e), (j), and (k) on the signs of s_w, $f''(k)$, and $m(k)$, respectively, the expression is seen to

[2] The intensive function is derived from the standard neoclassical constant returns function of K and L. In its gross form, by convention, it measures current production *before* the replacement of capital. The latter is assumed to depreciate fully during the production period.

[3] This condition is technically slightly stronger than strict quasi concavity. For the origin of this condition and elaboration of the technicality, see Donald Katzner, pp. 54; 210-11. I owe this point to Larry Selden.

[4] The second-order criterion in this paper will include $Z\equiv -u_{cc}+2u_{cc}u_c/u_x-u_{xx}u_c^2/u_x^2$. Since $u_x>0$ by assumption (c), letting $y_c=-1$, $y_x=u_c/u_x$ shows that assumption (b) implies $Z>0$.

[5] Government represents the interests of the young and not the old, because if the ethic is respected under all circumstances, government cannot affect, for better or worse, the economic situation of the old. It can thus be assumed that their participation in the political process is less substantial than that of the young whose future is not similarly insensitive to the actions of government.

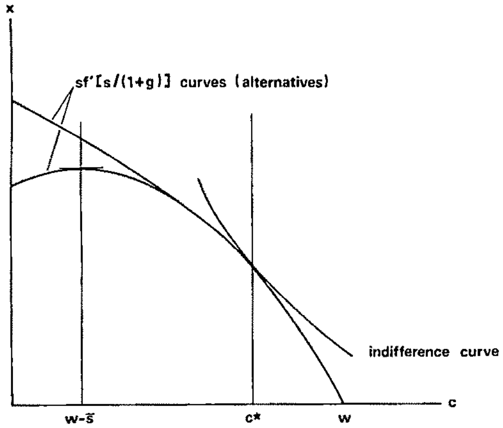


FIGURE 1

be positive, that is, k evolves monotonically over time.

The workings of this model can be seen more clearly on Figure 1, drawn for the representative young person, where the opportunities available to him through collective action are superimposed on his utility map.

Disposable income $w>0$ is marked off on the c axis. Consumption-saving possibilities as constrained by the distributional ethic are defined by the curve $x=sf'[s/(1+g)]$. Points on this curve satisfying $c\geqslant 0$, $x\geqslant 0$ constitute the opportunity set of the representative young person (ignoring available but definitely suboptimal points lying below the curve) and from this set the optimizing generation will choose the best point. That point is seen intuitively to be the one at which the $sf'[s/(1+g)]$ curve is tangent to an indifference curve. The tangency represents satisfaction of the first-order condition, since the numerical slope of the sf' curve is

$$(3)\quad f'[s/(1+g)]+sf''[s/(1+g)] \\ /(1+g)\equiv e[s/(1+g)]$$

Satisfaction of the second-order condition is obvious from the shapes of the curves.[6]

B. *Existence of the Optimum*

THEOREM 1: *Assumptions* (a), (d), (k), *and* (m) *are sufficient to guarantee the existence of a unique, interior solution to the generational optimization problem.* (*Here "interior" denotes satisfaction of $c\geqslant 0$, $x\geqslant 0$ with strict inequality.*)

PROOF:

The proof can most easily be demonstrated graphically on Figure 1. The set of feasible points is bounded below by $c\geqslant 0$, $x\geqslant 0$ and above by $c\leqslant w$ and, given the concavity of $sf'[s/(1+g)]$ established by Assumption (k), by $x\leqslant wf'[w/(1+g)]$ if sf' *does not* eventually decrease as s increases, or by $x\leqslant \bar{s}f'[\bar{s}/(1+g)]$, where $\bar{s}$ satisfies $f'[\bar{s}/(1+g)]+\bar{s}f''[\bar{s}/(1+g)]=0$, if sf' *does* eventually decrease. Given Assumption (m), the feasible set is closed. Therefore by the Weierstrass Theorem, the continuous function $u(c,x)$ has a maximum over this domain. If the maximum is at $c=0$, then $u_c=\infty$, and if at $x=0$, then $u_x=\infty$ by Assumption (d). Since $w>0$, these cannot be optimal positions. Thus the maximum is interior. By Assumption (k), sf' is concave. Given the convexity of the indifference curves there can be no more than one tangency.

C. *Comparison with Laissez-Faire*

By superimposing on a similar diagram (Figure 2) the *laissez-faire* consequences of the same initial conditions, it is possible to compare the initial period implications of the present model with those of *laissez-faire*.

At a point on the sf' curve with coordinates (c, sf'), there is saving per worker in the amount of $w-c$ or, equivalently, capital per next-period worker in the amount of $(w-c)/(1+g)$. The gross marginal product of this capital is given by the numerical slope of the straight line from (c, sf') to $(w,0)$ since $sf'/(w-c)=f'$.

If r is construed as parametric, the expression $x=(1+r)\ (w-c)$ defines a family of

[6] Note that the solution is identical to that which would obtain if parents put their children to work in a family cottage industry with binding custom dictating that the children be paid a marginal product wage. Capital's share of the product plus the value of the capital itself would be available for consumption by the retired parents. The size of the family industry during the subsequent period would then depend on the savings accumulated by the children who would be confronting an sf' curve and not a market interest rate.

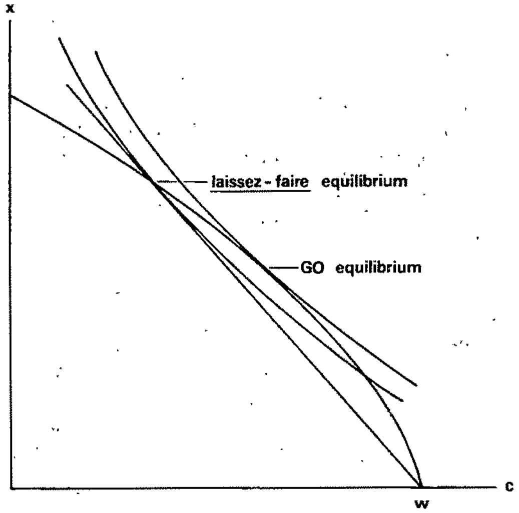


FIGURE 2

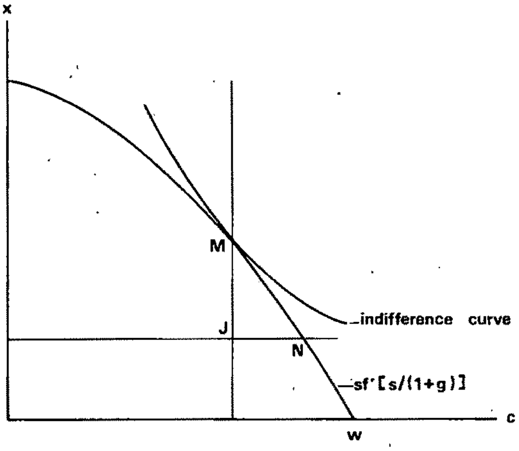


FIGURE 3

rays emanating from a given w. Each of these rays is tangent to some indifference curve; the locus of tangencies defines on offer (of capital) curve: $s(r)=w-c(r)$. In addition, each ray meets the sf' curve.[7] If $c_d(r)$ denotes the abscissa of this intersection, then $s_d(r)\equiv w-c_d(r)$ defines the capital demanded at r, that is, it satisfies $1+r=f'[s/(1+g)]$. Therefore the intersection of the offer curve with sf' represents the *laissez-faire* equilibrium. That such an equilibrium is unique when $1+g+f_t''c_r>0$ everywhere was established by Diamond.

THEOREM 2: *A generation optimizing collectively saves less than it would in a laissez-faire economy, given the same $w>0$.*

PROOF:

With reference to Figure 3, assume that the *laissez-faire* equilibrium (point N) is to the right of the monopoly equilibrium (point M on the sf' curve). By Assumption (k) on the concavity of sf', the line connecting N and $(w,0)$ has greater numerical slope than the sf' curve at M. However, the numerical slope of the indifference curve through N must be lower than that of the indifference curve through J, since $c_w<1$, which in turn must be lower than that of the indifference curve through M, since $c_w>0$. This is a contradiction. By (k) again, the tangent at M and the segment connecting M and $(w,0)$ have different slopes. A smooth indifference curve cannot be tangent to both at M. Thus N and M are not identical.

Two corollaries follow readily:

COROLLARY 1: *The function $k_t(k_{t-1})$ in the generational optimum economy lies beneath the function $k_t(k_{t-1})$ in the laissez-faire economy for all k_{t-1} such that $w(k_{t-1})>0$.*

PROOF:

Follows trivially from the theorem and $s\equiv k(1+g)$.

COROLLARY 2: *If both the laissez-faire and generational optimum economies have unique stable-state equilibria (not precluding the existence of unstable ones), designated k_D and k_M, respectively, then $k_D>k_M$.*

PROOF:

Follows trivially from Corollary 1. (See Figure 4.)

Two further consequences of the model will be of use in the following pages. First, given any common w_t, u_t is greater in the *GO*

[7] For sufficiently low r, the intersection will take place to the left of the ordinate. This is not depicted in the figures.

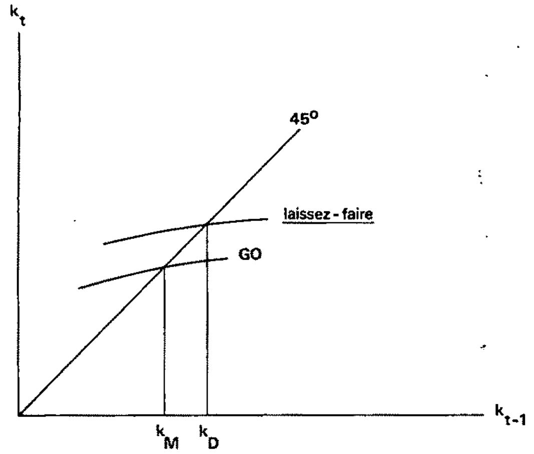


FIGURE 4

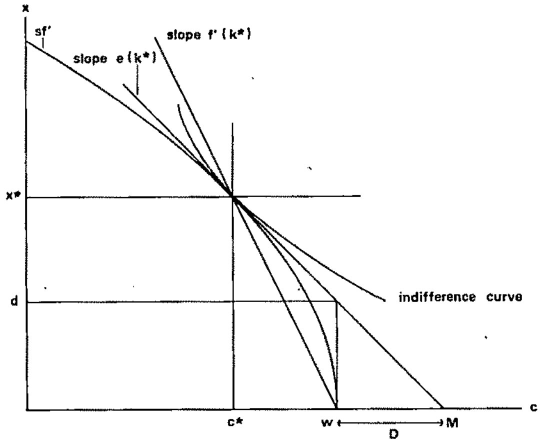


FIGURE 5

economy than in the *laissez-faire* economy. Obviously *GO* utility cannot be less than *laissez-faire* utility since the latter is available to a collectively optimizing generation. The utilities cannot be the same since, as is obvious from Figure 1, there is only one feasible point giving *GO* u_t and there $u_c/u_x = e(k) \neq f'$, that is, *laissez-faire* conditions cannot be satisfied. The conclusion follows.

Second, in Appendix B, comparative statics analysis shows that $du_t/dk_t = -k_t f''_t u_c > 0$ in the *GO* economy, that is, the utility enjoyed by generation *t* always increases with the capital it works, namely k_t. Intuitively it can be seen that higher *k* means higher *w* (since $w'(k) = -kf'' > 0$) which enlarges the opportunity set of the generation.

In summary, like a conventional monopolist choosing from the demand schedule facing him the point that suits him best, young worker/savers collectively evaluate the demand curve (for capital) facing them and select from it the optimal point,[8] certain that the integrity of their investment will be maintained and that marginal productivity interest will be paid thereon. The victims of these monopolists are members of the succeeding generation—and, less directly, all subsequent generations—who "inherit" less capital to work than under *laissez-faire* and thus receive to that extent a lower wage. These succeeding generations will, of course, themselves exploit posterity by exercising their own monopoly power in like manner.

D. *Implementation*

The solution to each generations's optimization problem can be implemented, and its monopoly power thus exercised, through an interest income tax-cum-transfer policy whereby all tax revenues collected are redistributed in lump sum form to the tax-paying generation. (See my dissertation where it is shown that a tax on saving coupled with a first-period redistribution can also serve to implement the optimum.)

Letting asterisks identify optimal values, the implementation is illustrated in Figure 5 where the optimum point (c^*, x^*) is depicted at the tangency of the *sf'* curve and the highest indifference curve touching it. As is evident from the figure, this optimum can be realized without intergenerational redistribution if unity plus the *perceived* (i.e., after-tax) interest rate equals $e(k^*)$—thus satisfying the first-order condition—and disposable income is (w, d) where *d* is a lump sum transfer dispensed in the same period taxes are paid.

The optimum tax rate T^* is equal to

$$(4)\quad ([f'(k^*)-1]-r^*)/[f'(k^*)-1]$$
$$\equiv -k^*f''(k^*)/[f'(k^*)-1]$$

[8] Note the similarity to optimum tariff theory as well.

The amount collected in taxes during the old period is then $s^*[f'(k^*)-1]T^* = -s^*k^*f''(k^*)$. By construction, an old period demogrant of d is perceived to be as valuable as $M-w=D$ so that a consumer's two-period "endowment" of (w, d) induces him to save s^* and select (c^*, x^*). It is seen that $d=x^*-s^*e(k^*)=s^*f'(k^*)-s^*e(k^*)=-s^*k^*f''(k^*)$ the amount collected in taxes. Thus what is taken from the old is returned; the government budget is balanced.[9]

II. Dynamic Implications

Clearly, a globally stable *GO* economy with unique equilibrium k will not in general evolve toward k_g, the Golden Rule capital-labor ratio (see Phelps; Diamond). This implies that if $k_0=k_g$, the economy will not in general sustain this level. Even in the special case where the economy does evolve toward or sustain k_g it does not evolve toward or sustain satisfaction of the *two-part* Golden Rule since in the steady state

$$(5) \qquad u_c/u_x=e(k_g)\neq f'(k_g)=1+g$$

where g is the biological rate of interest, that which gives in steady state the optimal distribution of a given product between generations.

I have demonstrated elsewhere that, for the *GO* economy, and for the Diamond *laissez-faire* economy, production function $f(k)=k^a$, $0<a<1$, and *CES* utility (no matter what the elasticity of substitution) are sufficient to guarantee global stability and the existence of a unique equilibrium k that is stable.[10]

With the existence established of certain economies which evolve toward some steady-state k from any $k>0$ under both *laissez-faire* and *GO* assumptions, it is possible to compare for these economies the normative implications of *laissez-faire* and *GO*.

It has already been shown that asymptotic equilibrium k must be less under the *GO* assumption than under *laissez-faire* (Corollary 2). This however implies nothing about asymptotic utility (designated u_D and u_M) in the two models. Although attempts to find necessary conditions for $u_D<u_M$, $u_M<u_D$ and $u_D=u_M$ have been unyielding, it can be shown at least that all three outcomes are possible under reasonable assumptions. This is demonstrated in Appendix C for the utility function $u(c,x)=b\cdot \log c+\log x$ and the production function $f(k)=k^a$.

It is at this point essential to determine whether either economy can produce a utility path that dominates the other's, given a common k_0. That the *laissez-faire* economy can never dominate is obvious from the fact that the first generation's utility (u_0) is always higher with monopoly power than without. However, since *GO* utility always starts out higher than *laissez-faire* utility and in some cases evolves to a higher equilibrium utility as well, it need only be shown that the two utility paths do not cross and recross each other in such a case to establish dominance of the *GO* path. Circumstances under which utility paths definitely do not cross twice are defined by Theorem 3 with the aid of the following lemma (proved elsewhere).

LEMMA: *In the laissez-faire economy, if $e(k)$ is positive for all positive k*[11] *and the k market is Walrasian stable, a generation's utility increases with the quantity of capital it works.* (See fn 10.)

THEOREM 3: *If $e(k)>0$ for all positive k and the k market is Walrasian stable, then for*

[9] The reader should by now be able to distinguish the present effort from its predecessors on one or more of the following counts: 1) In the *GO* economy there is no permanent government to implement an optimal plan, but rather a series of governments each with the mandate to promote the welfare of its (temporal) constituency alone. 2) The present concerns are explicitly dynamic; it is growth paths that are compared. (The papers by Diamond, Jerome Stein, and Toshihiro Ihori build on explicit dynamic processes, but their ultimate focus is on the steady state only.) 3) The governing ethic here is neither utilitarian nor Rawlsian. 4) The mechanism of government intervention is a tax on capital and not internal debt (Diamond, Stein, Ihori), intergenerational transfer (Ihori), or social capital (Stein).

[10] Formal proofs of all propositions unproved in this paper appear in my earlier discussion paper (1978b). They are also available upon request.

[11] That is, if the sf' curve's slope does not change sign (see Figure 1).

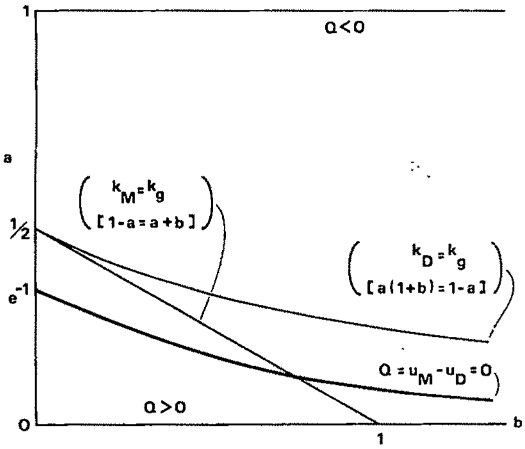


FIGURE 6

$k_M<k_0<k_D$, laissez-faire and GO utility paths do not cross twice.

PROOF:

If $k_0<k_D$, then *laissez-faire* k increases monotonically over time (see Diamond). Since by the lemma $e(k)>0$ for all positive k guarantees $du_t/dk_t>0$, utility increases over time as well. If $k_0=k_D$, k and u are constant over time. In the *GO* economy if $k_M<k_0$, k decreases monotonically. Since $du_t/dk_t>0$, u decreases likewise. If $k_0=k_M$, k and u are constant over time. Since the utility paths never move in the same direction over time, they can cross no more than once.

This result permits the following conclusion.

COROLLARY 3: *The GO utility path may dominate the laissez-faire utility path.*

PROOF:

Suppose $f(k)=k^a$ and $u(c,x)=b\cdot log\, c+log\, x$ and a and b are such that $u_D<u_M$. Then $e(k)=a^2k^{a-1}>0$ and so if $k_M<k_0<k_D$ and $u_D<u_M$, then the *GO* path dominates since it starts higher, evolves to a higher asymptote, and does not cross the *laissez-faire* path more than once by virtue of Theorem 3.

Through treatment of the case where $f(k)=k^a$ and $u(c,x)=b\cdot log\, c+log\, x$, it can be shown that overaccumulation in the *laissez-faire* economy can explain why that economy's utility path might be dominated by the *GO* path given the same initial k. (Overaccumulation means here that $k_g<k_D$.)

Let $Q\equiv u_M-u_D$. In the special case under consideration, there are certain values of a and b for which $Q=0$. (See Appendix C.) These are graphed on Figure 6. (Note that boundary points represent impermissible values of a and b. There is no upper bound on values of b.)

From (A4) of Appendix A,

$$(6)\quad k_D=\left[(1-a)/(1+g)(1+b)\right]^{1/(1-a)}$$

The Golden Rule k satisfies $ak_g^{a-1}=1+g$ or

$$(7)\qquad k_g=\left[a/(1+g)\right]^{1/(1-a)}$$

Thus all a and b satisfying $a(1+b)=1-a$ produce $k_D=k_g$. Since the two-part Golden Rule is satisfied, these points must lie in the $Q<0$ area and are so depicted.[12] Finally, from (A17) of Appendix B,

$$(8)\quad k_M=\left[a(1-a)/(1+g)(a+b)\right]^{1/(1-a)}$$

Thus all a and b satisfying $1-a=a+b$ produce $k_M=k_g$.

From Figure 6 it can be seen that overaccumulation in the *laissez-faire* economy need not result in $u_D<u_M$ if that overaccumulation is sufficiently small. Moreover, even when the *GO* economy evolves to k_g, it is still possible for an overaccumulating *laissez-faire* economy to give higher utility in the steady state. This is because

$$(9)\quad 1+r=e(k_M)<f'(k_M)=f'(k_g)\equiv 1+g$$

that is, the interest rate facing savers in the *GO* economy may be far from the biological optimum.

[12] For $k_D=k_g$, $(a+b)/(1+b)=1-a$ so that $Q=0$ implies $a^a=(1-a)^{1-a}$ by virtue of (A22). From the asymmetry of x^x between 0 and 1 (Figure 7) it is clear that only for the impermissible $a=1/2$, $b=0$ can this hold. Thus when $k_D=k_g$, Q can never be zero. As b approaches 0, a approaches $1/2>e^{-1}$ and $Q<0$. Thus if $k_D=k_g$, u_M is always inferior to u_D. This is, of course, just a special case of the well-known proposition that a self-sustaining Golden Rule steady state cannot be improved upon.

It has been shown that for certain functional forms and initial conditions, *every* generation may be better off in the *GO* economy than under *laissez-faire*. With each generation selfishly "playing" the demand curve for capital to the detriment of posterity, a utility stream unambiguously preferable to the *laissez-faire* sequence may in fact be produced. When such a sequence results, it is because government intervention in the saving decision has corrected, in part, a dynamic inefficiency.

III. Concluding Remarks

The above discussion has shown that an economy characterized by egoism and an intuitively appealing distributional ethic permits the intertemporal exercise of monopoly power. Under reasonable circumstances, a monopoly path may dominate the *laissez-faire* path that would evolve under indentical initial conditions.

While egoism and the ethical constraint are critical, the conclusions do not appear dependent on any other of the simplifying assumptions. If the consumption (or utility) of one's own offspring were an argument in his utility function, or if uncertainty as to life span existed, bequest motives would be introduced but they would not alter the model's underlying logic.[13] Continuous time would greatly complicate the mathematics, but there is no reason to expect fundamentally different conclusions. Finally, as shown in my 1978a, 1980 papers, with the population growth rate treated as an additional (or even the only) government choice variable, the model functions substantially as described above; *GO* assumptions reduce steady state k.

The ethical constraint has an intuitive appeal: no generation may take from any other. It might be criticized, though, on the grounds that it permits the exercise of monopoly power, believed by many to be inherently unjust, and thus helps rob defenseless unborn generations as surely as does the oft-decried plunder of natural resources. Critics must reckon, however, with each generation's ability to reassert monopoly power in its turn and with the irony that all generations may gain from monopoly. It is clear that there must be some constraint on government in a *GO*-type economy if society is to function. Not all clear, though, is whether a constraint with more intuitive ethical appeal than the present one can be formulated. Can it be ethically unacceptable for a generation to voluntarily tax itself? (Is it unacceptable for a cartel to tax, as a matter of policy, those members who exceed their quota?)

Appendix A

Steady-state values of c, x, and k must satisfy the following simultaneous equations in the Diamond economy:

$$c+k(1+g)-f(k)+kf'(k)=0 \tag{A1}$$

$$x-(1+g)kf'(k)=0 \tag{A2}$$

$$u_c-u_xf'(k)=0 \tag{A3}$$

For the special case where $f(k)=k^a$ and $u(c,x)=b\cdot log\,c+log\,x$, substitution permits explicit solutions:

$$k_D=[(1-a)/(1+b)(1+g)]^{1/(1-a)} \tag{A4}$$

$$c=b(1+g)k_D \tag{A5}$$

$$x=a(1+g)k_D^a \tag{A6}$$

Appendix B

The optimization problem in the *GO* economy is

$$\max_{c,x,k_t} u(c,x) \tag{A7}$$

subject to
$$c+k(1+g)-f(k_{t-1})+k_{t-1}f'(k_{t-1})=0$$
$$x-(1+g)k_tf'(k_t)=0$$

Solution of the problem yields:

$$u_c+\lambda_1=0 \tag{A8a}$$

$$u_x+\lambda_2=0 \tag{A8b}$$

$$(1+g)(\lambda_1-\lambda_2e(k_t))=0 \tag{A8c}$$

[13] Any generalized altruism, however, would be inconsistent with the important assumption of egoism.

$$\text{(A9)}\quad \begin{bmatrix} u_{cc} & u_{cx} & 0 & 1 & 0 \\ u_{cx} & u_{xx} & 0 & 0 & 1 \\ 0 & 0 & (1+g)u_x m(k_t) & 1+g & -(1+g)e(k_t) \\ 1 & 0 & 1+g & 0 & 0 \\ 0 & 1 & -(1+g)e(k_t) & 0 & 0 \end{bmatrix} \begin{bmatrix} dc \\ dx \\ dk_t \\ d\lambda_1 \\ d\lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -k_{t-1}f''_{t-1}dk_{t-1} \\ 0 \end{bmatrix},$$

and the bordered Hessian shown in (A9) above.

The first-order condition is $u_c-u_xe(k_t)=0$. The second-order condition for a maximum is $-Z+u_xm(k_t)/(1+g)<0$.

By substituting c_w, x_w, and c_r^* for their equivalents in terms of $u(c,x)$ derivatives,[14] the following can be derived:

$$\text{(A10)}\quad dc/dk_{t-1}=(-k_{t-1}f''_{t-1}) \times[(1+g)c_w+m(k_t)c_r^*]/[1+g+m(k_t)c_r^*]$$

$$\text{(A11)}\quad dx/dk_{t-1}=(-k_{t-1}f''_{t-1}) \times(1+g)x_w/[1+g+m(k_t)c_r^*]$$

$$\text{(A12)}\quad dk_t/dk_{t-1}=(-k_{t-1}f''_{t-1}) \times(1-c_w)/[1+g+m(k_t)c_r^*]$$

$$\text{(A13)}\quad du_t/dk_{t-1}=\frac{(-k_{t-1}f''_{t-1})}{[1+g+m(k_t)c_r^*]} \times[u_c(1+g)c_w+u_cm(k_t)c_r^*+u_x(1+g)x_w] =-k_{t-1}f''_{t-1}u_c$$

The steady state is defined by finding the solution for c, x and k in the following simultaneous equations:

$$\text{(A14)}\quad c+k(1+g)-f(k)+kf'(k)=0$$

$$\text{(A15)}\quad x-(1+g)kf'(k)=0$$

$$\text{(A16)}\quad u_c-u_xe(k)=0$$

For the special case where $f(k)=k^a$ and $u(c,x)=b\cdot log\, c+log\, x$, substitution into these equations shows that steady state k, c, and x can be solved for explicitly:

$$\text{(A17)}\quad k_M=[a(1-a)/(a+b)(1+g)]^{1/(1-a)}$$

$$\text{(A18)}\quad c=(b/a)(1+g)k_M$$

$$\text{(A19)}\quad x=a(1+g)k_M^a$$

APPENDIX C

It can be shown that, for an economy where $f(k)=k^a$ and $u(c,x)=b\cdot log\, c+log\, x$, steady-state utility in the *GO* economy (u_M) may exceed, equal, or be exceeded by steady-state utility under *laissez-faire* (u_D), given a common a, b, g. From equations (A5), (A6), (A18) and (A19), expressions for u_M and u_D can be derived. Letting Q denote u_M-u_D,

$$\text{(A20)}\quad Q=-b\cdot log\, a+(a+b)\cdot log(k_M/k_D)$$

Now substituting (A4) and (A17) for k_D and k_M

$$\text{(A21)}\quad Q=-b\cdot log\, a +\frac{(a+b)}{1-a}log\left[\frac{a(1+b)}{a+b}\right] =\left[\frac{1+b}{1-a}\right]\left[a\cdot log\, a-\left(\frac{a+b}{1+b}\right)log\left(\frac{a+b}{1+b}\right)\right]$$

This implies that

$$\text{(A22)}\quad Q=0 \text{ iff } a^a=[(a+b)/(1+b)]^{(a+b)/(1+b)}$$

Since $a<1$ and $b>0$, $a<(a+b)/(1+b)$. The problem then becomes determining whether for the function $y=x^x$ there can exist two different x values which give the same y, and if so, whether they can be equal to a and $(a+b)/(1+b)$ given the limitations

[14] These are standard results for the Fisher model. The derivations appear in my 1978a, b papers.

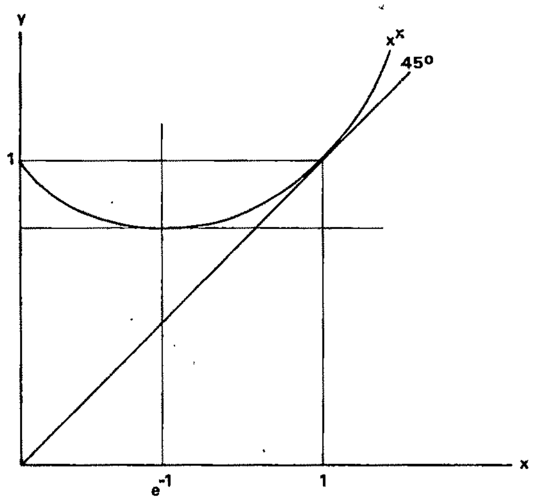


FIGURE 7

imposed by $0<a<1$ and $b>0$. Figure 7 depicts the behavior of x^x.

From the figure it can be seen that subject to the constraint $1>[(a+b)/(1+b)]>e^{-1}>a>0$, there is an interval of permissible a values for each of which there exists some permissible b value.

It will now be shown that Q may be positive or negative. Simple differentiation of the expression for Q yields:

$$\text{(A23)}\quad \frac{dQ}{da}=\frac{b}{1-a}+\frac{(1+b)}{(1-a)^2}\left[log\, a-log\left(\frac{a+b}{1+b}\right)\right]$$

$$\text{(A24)}\quad \frac{dQ}{db}=-\frac{1}{1+b}+\frac{1}{1-a}\left[a\, log\, a-log\left(\frac{a+b}{1+b}\right)\right]$$

Of interest are the values of these derivatives at $Q=0$. Substituting (A22) and simplifying yields:

$$\text{(A25)}\quad \frac{dQ}{da}=\frac{b}{1-a}\left[1+\frac{(1+b)}{(a+b)}\cdot log\, a\right]$$

$$\text{(A26)}\quad \frac{dQ}{db}=-\frac{1}{1+b}\left[1+log\left(\frac{a+b}{1+b}\right)\right]$$

Since $a<e^{-1}<(a+b)/(1+b)<1$, both derivatives are negative.

In summary, there exist values of a and b for which steady-state utility is identical in the two economies. Small decreases in a or b or both from any pair of these values result in higher utility for the *GO* economy, while small increases result in higher utility for the *laissez-faire* economy.

REFERENCES

P. A. Diamond, "National Debt in a Neoclassical Growth Model," *Amer. Econ. Rev.*, Dec. 1965, *55*, 1126–50.

Irving Fisher, *The Theory of Interest*, New York 1930.

T. Ihori, "The Golden Rule and the Role of Government in a Life Cycle Growth Model," *Amer. Econ. Rev.*, June 1978, *68*, 389–96.

Donald Katzner, *Static Demand Theory*, New York 1970.

L. D. Krohn, (1978a) "Intertemporal Monopoly Power: Models of Aggregative Growth with Government Intervention and an Ethical Constraint," unpublished doctoral dissertation, Columbia Univ. 1978.

______, (1978b) "The Generational Optimum Economy: Extracting Monopoly Gains from Posterity through the Income Tax," PSSP disc. paper no. 78-104, Oberlin College, Dec. 1978.

______, "Population Growth in the Generational Optimum Economy," document de travail #80-29, Laboratoire de recherche, Facultè des sciences de l'administration, Universitè Laval 1980.

Edmund S. Phelps, *Golden Rules of Economic Growth*, New York 1966.

P. A. Samuelson, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *J. Polit. Econ.*, Dec. 1958, *66*, 467–82.

J. L. Stein, "A Minimal Role of Government in Achieving Optimal Growth," *Economica*, May 1969, *34*, 139–50.

Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?

By ROBERT J. SHILLER*

A simple model that is commonly used to interpret movements in corporate common stock price indexes asserts that real stock prices equal the present value of rationally expected or optimally forecasted future real dividends discounted by a constant real discount rate. This valuation model (or variations on it in which the real discount rate is not constant but fairly stable) is often used by economists and market analysts alike as a plausible model to describe the behavior of aggregate market indexes and is viewed as providing a reasonable story to tell when people ask what accounts for a sudden movement in stock price indexes. Such movements are then attributed to "new information" about future dividends. I will refer to this model as the "efficient markets model" although it should be recognized that this name has also been applied to other models.

It has often been claimed in popular discussions that stock price indexes seem too "volatile," that is, that the movements in stock price indexes could not realistically be attributed to any objective new information, since movements in the price indexes seem to be "too big" relative to actual subsequent events. Recently, the notion that financial asset prices are too volatile to accord with efficient markets has received some econometric support in papers by Stephen LeRoy and Richard Porter on the stock market, and by myself on the bond market.

To illustrate graphically why it seems that stock prices are too volatile, I have plotted in Figure 1 a stock price index p_t with its *ex post* rational counterpart p_t^* (data set 1).[1] The stock price index p_t is the real Standard and Poor's Composite Stock Price Index (detrended by dividing by a factor proportional to the long-run exponential growth path) and p_t^* is the present discounted value of the actual subsequent real dividends (also as a proportion of the same long-run growth factor).[2] The analogous series for a modified Dow Jones Industrial Average appear in Figure 2 (data set 2). One is struck by the smoothness and stability of the *ex post* rational price series p_t^* when compared with the actual price series. This behavior of p^* is due to the fact that the present value relation relates p^* to a long-weighted moving average of dividends (with weights corresponding to discount factors) and moving averages tend to smooth the series averaged. Moreover, while real dividends did vary over this sample period, they did not vary long enough or far enough to cause major movements in p^*. For example, while one normally thinks of the Great Depression as a time when business was bad, real dividends were substantially below their long-run exponential growth path (i.e., 10–25 percent below the

*Associate professor, University of Pennsylvania, and research associate, National Bureau of Economic Research. I am grateful to Christine Amsler for research assistance, and to her as well as Benjamin Friedman, Irwin Friend, Sanford Grossman, Stephen LeRoy, Stephen Ross, and Jeremy Siegel for helpful comments. This research was supported by the National Bureau of Economic Research as part of the Research Project on the Changing Roles of Debt and Equity in Financing U.S. Capital Formation sponsored by the American Council of Life Insurance and by the National Science Foundation under grant SOC-7907561. The views expressed here are solely my own and do not necessarily represent the views of the supporting agencies.

[1] The stock price index may look unfamiliar because it is deflated by a price index, expressed as a proportion of the long-run growth path and only January figures are shown. One might note, for example, that the stock market decline of 1929–32 looks smaller than the recent decline. In real terms, it was. The January figures also miss both the 1929 peak and 1932 trough.

[2] The price and dividend series as a proportion of the long-run growth path are defined below at the beginning of Section I. Assumptions about public knowledge or lack of knowledge of the long-run growth path are important, as shall be discussed below. The series p^* is computed subject to an assumption about dividends after 1978. See text and Figure 3 below.

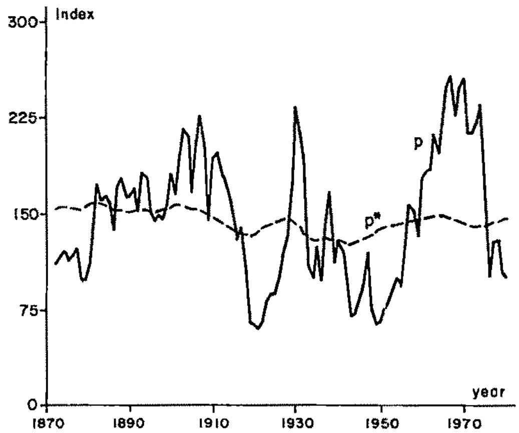


FIGURE 1
Note: Real Standard and Poor's Composite Stock Price Index (solid line p) and *ex post* rational price (dotted line p^*), 1871–1979, both detrended by dividing a long-run exponential growth factor. The variable p^* is the present value of actual subsequent real detrended dividends, subject to an assumption about the present value in 1979 of dividends thereafter. Data are from Data Set 1, Appendix.

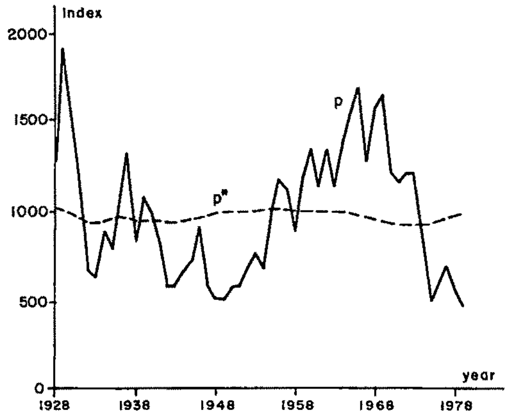


FIGURE 2
Note: Real modified Dow Jones Industrial Average (solid line p) and *ex post* rational price (dotted line p^*), 1928-1979, both detrended by dividing by a long-run exponential growth factor. The variable p^* is the present value of actual subsequent real detrended dividends, subject to an assumption about the present value in 1979 of dividends thereafter. Data are from Data Set 2, Appendix.

growth path for the Standard and Poor's series, 16–38 percent below the growth path for the Dow Series) only for a few depression years: 1933, 1934, 1935, and 1938. The moving average which determines p^* will smooth out such short-run fluctuations. Clearly the stock market decline beginning in 1929 and ending in 1932 could not be rationalized in terms of subsequent dividends! Nor could it be rationalized in terms of subsequent earnings, since earnings are relevant in this model only as indicators of later dividends. Of course, the efficient markets model does not say $p=p^*$. Might one still suppose that this kind of stock market crash was a rational mistake, a forecast error that rational people might make? This paper will explore here the notion that the very volatility of p (i.e., the tendency of big movements in p to occur again and again) implies that the answer is no.

To give an idea of the kind of volatility comparisons that will be made here, let us consider at this point the simplest inequality which puts limits on one measure of volatility: the standard deviation of p. The efficient markets model can be described as asserting that $p_t=E_t(p_t^*)$, i.e., p_t is the mathematical expectation conditional on all information available at time t of p_t^*. In other words, p_t is the optimal forecast of p_t^*. One can define the forecast error as $u_t=p_t^*-p_t$. A fundamental principle of optimal forecasts is that the forecast error u_t must be uncorrelated with the forecast; that is, the covariance between p_t and u_t must be zero. If a forecast error showed a consistent correlation with the forecast itself, then that would in itself imply that the forecast could be improved. Mathematically, it can be shown from the theory of conditional expectations that u_t must be uncorrelated with p_t.

If one uses the principle from elementary statistics that the variance of the sum of two uncorrelated variables is the sum of their variances, one then has $var(p^*)=var(u)+var(p)$. Since variances cannot be negative, this means $var(p)\leqslant var(p^*)$ or, converting to more easily interpreted standard deviations,

$$(1) \qquad \sigma(p)\leqslant\sigma(p^*)$$

This inequality (employed before in the

papers by LeRoy and Porter and myself) is violated dramatically by the data in Figures 1 and 2 as is immediately obvious in looking at the figures.[3]

This paper will develop the efficient markets model in Section I to clarify some theoretical questions that may arise in connection with the inequality (1) and some similar inequalities will be derived that put limits on the standard deviation of the innovation in price and the standard deviation of the change in price. The model is restated in innovation form which allows better understanding of the limits on stock price volatility imposed by the model. In particular, this will enable us to see (Section II) that the standard deviation of Δp is highest when information about dividends is revealed smoothly and that if information is revealed in big lumps occasionally the price series may have higher kurtosis (fatter tails) but will have *lower* variance. The notion expressed by some that earnings rather than dividend data should be used is discussed in Section III, and a way of assessing the importance of time variation in real discount rates is shown in Section IV. The inequalities are compared with the data in Section V.

This paper takes as its starting point the approach I used earlier (1979) which showed evidence suggesting that long-term bond yields are too volatile to accord with simple expectations models of the term structure of interest rates.[4] In that paper, it was shown how restrictions implied by efficient markets on the cross-covariance function of short-term and long-term interest rates imply inequality restrictions on the spectra·of the long-term interest rate series which characterize the smoothness that the long rate should display. In this paper, analogous implications are derived for the volatility of stock prices, although here a simpler and more intuitively appealing discussion of the model in terms of its innovation representation is used. This paper also has benefited from the earlier discussion by LeRoy and Porter which independently derived some restrictions on security price volatility implied by the efficient markets model and concluded that common stock prices are too volatile to accord with the model. They applied a methodology in some ways similar to that used here to study a stock price index and individual stocks in a sample period starting after World War II.

It is somewhat inaccurate to say that this paper attempts to contradict the extensive literature of efficient markets (as, for example, Paul Cootner's volume on the random character of stock prices, or Eugene Fama's survey).[5] Most of this literature really examines different properties of security prices. Very little of the efficient markets literature bears directly on the characteristic feature of the model considered here: that expected *real* returns for the aggregate stock market are constant through time (or approximately so). Much of the literature on efficient markets concerns the investigation of nominal "profit opportunities" (variously defined) and whether transactions costs prohibit their exploitation. Of course, if real stock prices are "too volatile" as it is defined here, then there may well be a sort of real profit opportunity. Time variation in expected real interest rates does not itself imply that any

[3] Some people will object to this derivation of (1) and say that one might as well have said that $E_t(p_t)=p_t^*$, i.e., that forecasts are correct "on average," which would lead to a reversal of the inequality (1). This objection stems, however, from a misinterpretation of conditional expectations. The subscript t on the expectations operator E means "taking as given (i.e., nonrandom) all variables known at time t." Clearly, p_t is known at time t and p_t^* is not. In practical terms, if a forecaster gives as his forecast anything other than $E_t(p_t^*)$, then high forecast is not optimal in the sense of expected squared forecast error. If he gives a forecast which equals $E_t(p_t^*)$ only on average, then he is adding random noise to the optimal forecast. The amount of noise apparent in Figures 1 or 2 is extraordinary. Imagine what we would think of our local weather forecaster if, say, actual local temperatures followed the dotted line and his forecasts followed the solid line!

[4] This analysis was extended to yields on preferred stocks by Christine Amsler.

[5] It should not be inferred that the literature on efficient markets uniformly supports the notion of efficiency put forth there, for example, that no assets are dominated or that no trading rule dominates a buy and hold strategy, (for recent papers see S. Basu; Franco Modigliani and Richard Cohn; William Brainard, John Shoven and Lawrence Weiss; and the papers in the symposium on market efficiency edited by Michael Jensen).

trading rule dominates a buy and hold strategy, but really large variations in expected returns might seem to suggest that such a trading rule exists. This paper does not investigate this, or whether transactions costs prohibit its exploitation. This paper is concerned, however, instead with a more interesting (from an economic standpoint) question: what accounts for movements in real stock prices and can they be explained by new information about subsequent real dividends? If the model fails due to excessive volatility, then we will have seen a new characterization of how the simple model fails. The characterization is not equivalent to other characterizations of its failure, such as that one-period holding returns are forecastable, or that stocks have not been good inflation hedges recently.

The volatility comparisons that will be made here have the advantage that they are insensitive to misalignment of price and dividend series, as may happen with earlier data when collection procedures were not ideal. The tests are also not affected by the practice, in the construction of stock price and dividend indexes, of dropping certain stocks from the sample occasionally and replacing them with other stocks, so long as the volatility of the series is not misstated. These comparisons are thus well suited to existing long-term data in stock price averages. The robustness that the volatility comparisons have, coupled with their simplicity, may account for their popularity in casual discourse.

I. The Simple Efficient Markets Model

According to the simple efficient markets model, the real price P_t of a share at the beginning of the time period t is given by

$$(2) \qquad P_t = \sum_{k=0}^{\infty} \gamma^{k+1} E_t D_{t+k} \qquad 0<\gamma<1$$

where D_t is the real dividend paid at (let us say, the end of) time t, E_t denotes mathematical expectation conditional on information available at time t, and γ is the constant real discount factor. I define the constant real interest rate r so that $\gamma = 1/(1+r)$. Information at time t includes P_t and D_t and their lagged values, and will generally include other variables as well.

The one-period holding return $H_t \equiv (\Delta P_{t+1} + D_t)/P_t$ is the return from buying the stock at time t and selling it at time $t+1$. The first term in the numerator is the capital gain, the second term is the dividend received at the end of time t. They are divided by P_t to provide a rate of return. The model (2) has the property that $E_t(H_t) = r$.

The model (2) can be restated in terms of series as a proportion of the long-run growth factor: $p_t = P_t/\lambda^{t-T}$, $d_t = D_t/\lambda^{t+1-T}$ where the growth factor is $\lambda^{t-T} = (1+g)^{t-T}$, g is the rate of growth, and T is the base year. Dividing (2) by λ^{t-T} and substituting one finds[6]

$$(3) \qquad p_t = \sum_{k=0}^{\infty} (\lambda\gamma)^{k+1} E_t d_{t+k}$$

$$= \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} E_t d_{t+k}$$

The growth rate g must be less than the discount rate r if (2) is to give a finite price, and hence $\bar{\gamma} \equiv \lambda\gamma < 1$, and defining $\bar{r}$ by $\bar{\gamma} \equiv 1/(1+\bar{r})$, the discount rate appropriate for the p_t and d_t series is $\bar{r} > 0$. This discount rate $\bar{r}$ is, it turns out, just the mean dividend divided by the mean price, i.e, $\bar{r} = E(d)/E(p)$.[7]

[6] No assumptions are introduced in going from (2) to (3), since (3) is just an algebraic transformation of (2). I shall, however, introduce the assumption that d_t is jointly stationary with information, which means that the (unconditional) covariance between d_t and z_{t-k}, where z_t is any information variable (which might be d_t itself or p_t), depends only on k, not t. It follows that we can write expressions like $var(p)$ without a time subscript. In contrast, a realization of the random variable the *conditional* expectation $E_t(d_{t+k})$ is a function of time since it depends on information at time t. Some stationarity assumption is necessary if we are to proceed with any statistical analysis.

[7] Taking unconditional expectations of both sides of (3) we find

$$E(p) = \frac{\bar{\gamma}}{1-\bar{\gamma}} E(d)$$

using $\bar{\gamma} = 1/1 + \bar{r}$ and solving we find $\bar{r} = E(d)/E(p)$.

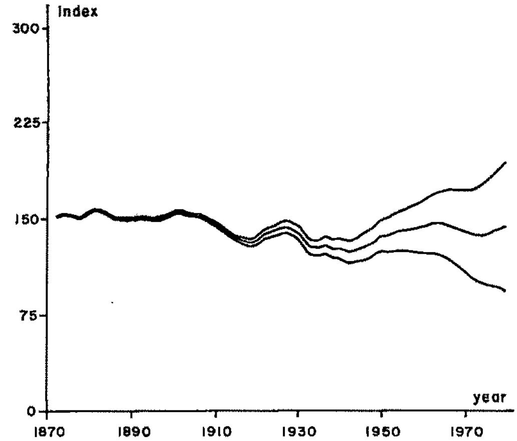


FIGURE 3

Note: Alternative measures of the *ex post* rational price p^*, obtained by alternative assumptions about the present value in 1979 of dividends thereafter. The middle curve is the p^* series plotted in Figure 1. The series are computed recursively from terminal conditions using dividend series d of Data Set 1.

TABLE 1—DEFINITIONS OF PRINCIPAL SYMBOLS

γ = real discount factor for series before detrending; $\gamma = 1/(1+r)$
$\bar{\gamma}$ = real discount factor for detrended series; $\bar{\gamma} \equiv \lambda\gamma$
D_t = real dividend accruing to stock index (before detrending)
d_t = real detrended dividend; $d_t \equiv D_t/\lambda^{t+1-T}$
Δ = first difference operator $\Delta x_t \equiv x_t - x_{t-1}$
δ_t = innovation operator; $\delta_t x_{t+k} \equiv E_t x_{t+k} - E_{t-1} x_{t+k}$; $\delta x \equiv \delta_t x_t$
E = unconditional mathematical expectations operator. $E(x)$ is the true (population) mean of x.
E_t = mathematical expectations operator conditional on information at time t; $E_t x_t \equiv E(x_t | I_t)$ where I_t is the vector of information variables known at time t.
λ = trend factor for price and dividend series; $\lambda \equiv 1+g$ where g is the long-run growth rate of price and dividends.
P_t = real stock price index (before detrending)
p_t = real detrended stock price index; $p_t = P_t/\lambda^{t-T}$
p_t^* = *ex post* rational stock price index (expression 4)
r = one-period real discount rate for series before detrending
$\bar{r}$ = real discount rate for detrended series; $\bar{r} = (1-\bar{\gamma})/\bar{\gamma}$
$\bar{r}_2$ = two-period real discount rate for detrended series; $\bar{r}_2 = (1+\bar{r})^2 - 1$
t = time (year)
T = base year for detrending and for wholesale price index; $p_T = P_T$ = nominal stock price index at time T

We may also write the model as noted above in terms of the *ex post* rational price series p_t^* (analogous to the *ex post* rational interest rate series that Jeremy Siegel and I used to study the Fisher effect, or that I used to study the expectations theory of the term structure). That is, p_t^* is the present value of actual subsequent dividends:

$$(4) \qquad p_t = E_t(p_t^*)$$

$$\text{where} \qquad p_t^* = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} d_{t+k}$$

Since the summation extends to infinity, we never observe p_t^* without some error. However, with a long enough dividend series we may observe an approximate p_t^*. If we choose an arbitrary value for the terminal value of p_t^* (in Figures 1 and 2, p^* for 1979 was set at the average detrended real price over the sample) then we may determine p_t^* recursively by $p_t^* = \bar{\gamma}(p_{t+1}^* + d_t)$ working backward from the terminal date. As we move back from the terminal date, the importance of the terminal value chosen declines. In data set (1) as shown in Figure 1, $\bar{\gamma}$ is .954 and $\bar{\gamma}^{108} = .0063$ so that at the beginning of the sample the terminal value chosen has a negligible weight in the determination of p_t^*. If we had chosen a different terminal condition, the result would be to add or subtract an exponential trend from the p^* shown in Figure 1. This is shown graphically in Figure 3, in which p^* is shown computed from alternative terminal values. Since the only thing we need know to compute p^* about dividends after 1978 is p^* for 1979, it does not matter whether dividends are "smooth" or not after 1978. Thus, Figure 3 represents our uncertainty about p^*.

There is yet another way to write the model, which will be useful in the analysis which follows. For this purpose, it is convenient to adopt notation for the innovation in a variable. Let us define the innovation operator $\delta_t \equiv E_t - E_{t-1}$ where E_t is the conditional expectations operator. Then for any variable X_t the term $\delta_t X_{t+k}$ equals $E_t X_{t+k} - E_{t-1} X_{t+k}$ which is the change in the conditional expectation of X_{t+k} that is made in response to new information arriving between $t-1$ and t. The time subscript t may be dropped so that δX_k denotes $\delta_t X_{t+k}$ and

δX denotes δX_0 or $\delta_t X_t$. Since conditional expectations operators satisfy $E_j E_k = E_{min(j,k)}$ it follows that $E_{t-m}\delta_t X_{t+k} = E_{t-m}(E_t X_{t+k} - E_{t-1} X_{t+k}) = E_{t-m} X_{t+k} - E_{t-m} X_{t+k} = 0$, $m \geqslant 0$. This means that $\delta_t X_{t+k}$ must be uncorrelated for all k with all information known at time $t-1$ and must, since lagged innovations are information at time t, be uncorrelated with $\delta_{t'} X_{t+j}$, $t' < t$, all j, i.e., innovations in variables are serially uncorrelated.

The model implies that the innovation in price $\delta_t p_t$ is observable. Since (3) can be written $p_t = \bar{\gamma}(d_t + E_t p_{t+1})$, we know, solving, that $E_t p_{t+1} = p_t/\bar{\gamma} - d_t$. Hence $\delta_t p_t \equiv E_t p_t - E_{t-1} p_t = p_t + d_{t-1} - p_{t-1}/\bar{\gamma} = \Delta p_t + d_{t-1} - \bar{r} p_{t-1}$. The variable which we call $\delta_t p_t$ (or just δp) is the variable which Clive Granger and Paul Samuelson emphasized should, in contrast to $\Delta p_t \equiv p_t - p_{t-1}$, by efficient markets, be unforecastable. In practice, with our data, $\delta_t p_t$ so measured will approximately equal Δp_t.

The model also implies that the innovation in price is related to the innovations in dividends by

$$\delta_t p_t = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \delta_t d_{t+k} \tag{5}$$

This expression is identical to (3) except that δ_t replaces E_t. Unfortunately, while $\delta_t p_t$ is observable in this model, the $\delta_t d_{t+k}$ terms are not directly observable, that is, we do not know when the public gets information about a particular dividend. Thus, in deriving inequalities below, one is obliged to assume the "worst possible" pattern of information accrual.

Expressions (2)–(5) constitute four different representations of the same efficient markets model. Expressions (4) and (5) are particularly useful for deriving our inequalities on measures of volatility. We have already used (4) to derive the limit (1) on the standard deviation of p given the standard deviation of p^*, and we will use (5) to derive a limit on the standard deviation of δp given the standard deviation of d.

One issue that relates to the derivation of (1) can now be clarified. The inequality (1) was derived using the assumption that the forecast error $u_t = p_t^* - p_t$ is uncorrelated with p_t. However, the forecast error u_t is not serially uncorrelated. It is uncorrelated with all information known at time t, but the lagged forecast error u_{t-1} is not known at time t since p_{t-1}^* is not discovered at time t. In fact, $u_t = \Sigma_{k=1}^{\infty} \bar{\gamma}^k \delta_{t+k} p_{t+k}$, as can be seen by substituting the expressions for p_t and p_t^* from (3) and (4) into $u_t = p_t^* - p_t$, and rearranging. Since the series $\delta_t p_t$ is serially uncorrelated, u_t has first-order autoregressive serial correlation.[8] For this reason, it is inappropriate to test the model by regressing $p_t^* - p_t$ on variables known at time t and using the ordinary t-statistics of the coefficients of these variables. However, a generalized least squares transformation of the variables would yield an appropriate regression test. We might thus regress the transformed variable $u_t - \bar{\gamma} u_{t+1}$ on variables known at time t. Since $u_t - \bar{\gamma} u_{t+1} = \bar{\gamma}\delta_{t+1} p_{t+1}$, this amounts to testing whether the innovation in price can be forecasted. I will perform and discuss such regression tests in Section V below.

To find a limit on the standard deviation of δp for a given standard deviation of d_t, first note that d_t equals its unconditional expectation plus the sum of its innovations:

$$d_t = E(d) + \sum_{k=0}^{\infty} \delta_{t-k} d_t \tag{6}$$

If we regard $E(d)$ as $E_{-\infty}(d_t)$, then this expression is just a tautology. It tells us, though, that d_t $t = 0, 1, 2, \ldots$ are just different linear combinations of the same innovations in dividends that enter into the linear combination in (5) which determine $\delta_t p_t$ $t = 0, 1, 2, \ldots$. We can thus ask how large *var* (δp) might be for given *var*(d). Since innovations are serially uncorrelated, we know from (6) that the variance of the sum is

[8] It follows that $var(u) = var(\delta p)/(1-\bar{\gamma}^2)$ as LeRoy and Porter noted. They base their volatility tests on our inequality (1) (which they call theorem 2) and an equality restriction $\sigma^2(p) + \sigma^2(\delta p)/(1-\bar{\gamma}^2) = \sigma^2(p^*)$ (their theorem 3). They found that, with postwar Standard and Poor earnings data, both relations were violated by sample statistics.

the sum of the variances:

$$(7)\qquad var(d)=\sum_{k=0}^{\infty} var(\delta d_k)=\sum_{k=0}^{\infty}\sigma_k^2$$

Our assumption of stationarity for d_t implies that $var(\delta_{t-k}d_t)\equiv var(\delta d_k)\equiv\sigma_k^2$ is independent of t.

In expression (5) we have no information that the variance of the sum is the sum of the variances since all the innovations are time t innovations, which may be correlated. In fact, for given $\sigma_0^2, \sigma_1^2, \ldots,$ the maximum variance of the sum in (5) occurs when the elements in the sum are perfectly positively correlated. This means then that so long as $var(\delta d)\neq 0$, $\delta_t d_{t+k}=a_k\delta_t d_t$, where $a_k=\sigma_k/\sigma_0$. Substituting this into (6) implies

$$(8)\qquad \hat{d}_t=\sum_{k=0}^{\infty} a_k\varepsilon_{t-k}$$

where a hat denotes a variable minus its mean: $\hat{d}_t\equiv d_t-E(d)$ and $\varepsilon_t\equiv\delta_t d_t$. Thus, if $var(\delta p)$ is to be maximized for given $\sigma_0^2, \sigma_1^2, \ldots,$ the dividend process must be a moving average process in terms of its own innovations.[9] I have thus shown, rather than assumed, that if the variance of δp is to be maximized, the forecast of d_{t+k} will have the usual ARIMA form as in the forecast popularized by Box and Jenkins.

We can now find the maximum possible variance for δp for given variance of d. Since the innovations in (5) are perfectly positively correlated, $var(\delta p)=(\Sigma_{k=0}^{\infty}\bar{\gamma}^{k+1}\sigma_k)^2$. To maximize this subject to the constraint $var(d)=\Sigma_{k=0}^{\infty}\sigma_k^2$ with respect to $\sigma_0, \sigma_1, \ldots,$ one may set up the Lagrangean:

$$(9)\qquad L=\left(\sum_{k=0}^{\infty}\bar{\gamma}^{k+1}\sigma_k\right)^2+\nu\left(var(d)-\sum_{k=0}^{\infty}\sigma_k^2\right)$$

where ν is the Lagrangean multiplier. The first-order conditions for $\sigma_j, j=0,\ldots\infty$ are

$$(10)\qquad \frac{\partial L}{\partial\sigma_j}=2\left(\sum_{k=0}^{\infty}\bar{\gamma}^{k+1}\sigma_k\right)\bar{\gamma}^{j+1}-2\nu\sigma_j=0$$

which in turn means that σ_j is proportional to $\bar{\gamma}^j$. The second-order conditions for a maximum are satisfied, and the maximum can be viewed as a tangency of an isoquant for $var(\delta p)$, which is a hyperplane in $\sigma_0, \sigma_1, \sigma_2, \ldots$ space, with the hypersphere represented by the constraint. At the maximum $\sigma_k^2=(1-\bar{\gamma}^2)var(d)\bar{\gamma}^{2k}$ and $var(\delta p)=\bar{\gamma}^2 var(d)/(1-\bar{\gamma}^2)$ and so, converting to standard deviations for ease of interpretation, we have

$$(11)\qquad \sigma(\delta p)\leqslant\sigma(d)/\sqrt{\bar{r}_2}$$

$$\text{where}\qquad \bar{r}_2=(1+\bar{r})^2-1$$

Here, $\bar{r}_2$ is the two-period interest rate, which is roughly twice the one-period rate. The maximum occurs, then, when d_t is a first-order autoregressive process, $\hat{d}_t=\bar{\gamma}\hat{d}_{t-1}+\varepsilon_t$, and $E_t\hat{d}_{t+k}=\bar{\gamma}^k\hat{d}_t$, where $\hat{d}\equiv d-E(d)$ as before.

The variance of the innovation in price is thus maximized when information about dividends is revealed in a smooth fashion so that the standard deviation of the new information at time t about a future dividend d_{t+k} is proportional to its weight in the present value formula in the model (5). In contrast, suppose all dividends somehow became known years before they were paid. Then the innovations in dividends would be so heavily discounted in (5) that they would contribute little to the standard deviation of the innovation in price. Alternatively, suppose nothing were known about dividends until the year they are paid. Here, although the innovation would not be heavily discounted in (5), the impact of the innovation would be confined to only one term in (5), and the standard deviation in the innovation in price would be limited to the standard deviation in the single dividend.

Other inequalities analogous to (11) can also be derived in the same way. For exam-

[9] Of course, all indeterministic stationary processes can be given linear moving average representations, as Hermann Wold showed. However, it does not follow that the process can be given a moving average representation in terms of its own innovations. The true process may be generated nonlinearly or other information besides its own lagged values may be used in forecasting. These will generally result in a less than perfect correlation of the terms in (5).

ple, we can put an upper bound to the standard deviation of the change in price (rather than the innovation in price) for given standard deviation in dividend. The only difference induced in the above procedure is that Δp_t is a different linear combination of innovations in dividends. Using the fact that $\Delta p_t = \delta_t p_t + \bar{r} p_{t-1} - d_{t-1}$ we find

$$\Delta p_t = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \delta_t d_{t+k} + \bar{r} \sum_{j=1}^{\infty} \delta_{t-j} \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} d_{t+k-1} - \sum_{j=1}^{\infty} \delta_{t-j} d_{t-1} \tag{12}$$

As above, the maximization of the variance of δp for given variance of d requires that the time t innovations in d be perfectly correlated (innovations at different times are necessarily uncorrelated) so that again the dividend process must be forecasted as an ARIMA process. However, the parameters of the ARIMA process for d which maximize the variance of Δp will be different. One finds, after maximizing the Lagrangean expression (analogous to (9)) an inequality slightly different from (11),

$$\sigma(\Delta p) \leqslant \sigma(d)/\sqrt{2\bar{r}} \tag{13}$$

The upper bound is attained if the optimal dividend forecast is first-order autoregressive, but with an autoregressive coefficient slightly different from that which induced the upper bound to (11). The upper bound to (13) is attained if $\hat{d}_t = (1-\bar{r})\hat{d}_{t-1} + \varepsilon_t$ and $E_t d_{t+k} = (1-\bar{r})^k \hat{d}_t$, where, as before, $\hat{d}_t \equiv d_t - E(d)$.

II. High Kurtosis and Infrequent Important Breaks in Information

It has been repeatedly noted that stock price change distributions show high kurtosis or "fat tails." This means that, if one looks at a time-series of observations on δp or Δp, one sees long stretches of time when their (absolute) values are all rather small and then an occasional extremely large (absolute) value. This phenomenon is commonly attributed to a tendency for new information to come in big lumps infrequently. There seems to be a common presumption that this information lumping might cause stock price changes to have high or infinite variance, which would seem to contradict the conclusion in the preceding section that the variance of price is limited and is maximized if forecasts have a simple autoregressive structure.

High sample kurtosis does not indicate infinite variance if we do not assume, as did Fama (1965) and others, that price changes are drawn from the stable Paretian class of distributions.[10] The model does not suggest that price changes have a distribution in this class. The model instead suggests that the existence of moments for the price series is implied by the existence of moments for the dividends series.

As long as d is jointly stationary with information and has a finite variance, then p, p^*, δp, and Δp will be stationary and have a finite variance.[11] If d is normally distributed, however, it does not follow that the price variables will be normally distributed. In fact, they may yet show high kurtosis.

To see this possibility, suppose the dividends are serially independent and identically normally distributed. The kurtosis of the price series is defined by $K = E(\hat{p})^4/(E(\hat{p})^2)^2$, where $p \equiv \hat{p} - E(p)$. Suppose, as an example, that with a probability of $1/n$

[10] The empirical fact about the unconditional distribution of stock price changes in not that they have infinite variance (which can never be demonstrated with any finite sample), but that they have high kurtosis in the sample.

[11] With any stationary process X_t, the existence of a finite $var(X_t)$ implies, by Schwartz's inequality, a finite value of $cov(X_t, X_{t+k})$ for any k, and hence the entire autocovariance function of X_t, and the spectrum, exists. Moreover, the variance of $E_t(X_t)$ must also be finite, since the variance of X equals the variance of $E_t(X_t)$ plus the variance of the forecast error. While we may regard real dividends as having finite variance, innovations in dividends may show high kurtosis. The residuals in a second-order autoregression for d_t have a studentized range of 6.29 for the Standard and Poor series and 5.37 for the Dow series. According to the David-Hartley-Pearson test, normality can be rejected at the 5 percent level (but not at the 1 percent level) with a one-tailed test for both data sets.

the public is told d_t at the beginning of time t, but with probability $(n-1)/n$ has no information about current or future dividends.[12] In time periods when they are told d_t, $\hat{p}_t$ equals $\bar{\gamma}\hat{d}_t$, otherwise $\hat{p}_t=0$. Then $E(\hat{p}_t^4)=E((\bar{\gamma}\hat{d}_t)^4)/n$ and $E(\hat{p}_t^2)=E((\bar{\gamma}\hat{d}_t)^2)/n$ so that kurtosis equals $nE(\bar{\gamma}\hat{d}_t)^4)/E((\bar{\gamma}\hat{d}_t)^2)$ which equals n times the kurtosis of the normal distribution. Hence, by choosing n high enough one can achieve an arbitrarily high kurtosis, and yet the variance of price will always exist. Moreover, the distribution of $\hat{p}_t$ conditional on the information that the dividend has been revealed is also normal, in spite of high kurtosis of the unconditional distribution.

If information is revealed in big lumps occasionally (so as to induce high kurtosis as suggested in the above example) $var(\delta p)$ or $var(\Delta p)$ are not especially large. The variance loses more from the long interval of time when information is not revealed than it gains from the infrequent events when it is. The highest possible variance for given variance of d indeed comes when information is revealed smoothly as noted in the previous section. In the above example, where information about dividends is revealed one time in n, $\sigma(\delta p)=\bar{\gamma}n^{1/2}\sigma(d)$ and $\sigma(\Delta p)=\bar{\gamma}(2/n)^{1/2}\sigma(d)$. The values of $\sigma(\delta p)$ and $\sigma(\Delta p)$ implied by this example are for all n strictly below the upper bounds of the inequalities (11) and (13).[13]

III. Dividends or Earnings?

It has been argued that the model (2) does not capture what is generally meant by efficient markets, and that the model should be replaced by a model which makes price the present value of expected earnings rather than dividends. In the model (2) earnings may be relevant to the pricing of shares but only insofar as earnings are indicators of future dividends. Earnings are thus no different from any other economic variable which may indicate future dividends. The model (2) is consistent with the usual notion in finance that individuals are concerned with returns, that is, capital gains plus dividends. The model implies that expected total returns are constant and that the capital gains component of returns is just a reflection of information about future dividends. Earnings, in contrast, are statistics conceived by accountants which are supposed to provide an indicator of how well a company is doing, and there is a great deal of latitude for the definition of earnings, as the recent literature on inflation accounting will attest.

There is no reason why price per share ought to be the present value of expected earnings per share if some earnings are retained. In fact, as Merton Miller and Franco Modigliani argued, such a present value formula would entail a fundamental sort of double counting. It is incorrect to include in the present value formula both earnings at time t and the later earnings that accrue when time t earnings are reinvested.[14] Miller and Modigliani showed a formula by which price might be regarded as the present value of earnings corrected for investments, but that formula can be shown, using an accounting identity to be identical to (2).

Some people seem to feel that one cannot claim price as present value of expected dividends since firms routinely pay out only a fraction of earnings and also attempt somewhat to stabilize dividends. They are right in the case where firms paid out *no* dividends, for then the price p_t would have to grow at the discount rate $\bar{r}$, and the model (2) would not be the solution to the difference equation implied by the condition $E_t(H_t)=r$. On the other hand, if firms pay out a fraction of dividends or smooth short-run fluctuations in dividends, then the price of the firm will grow at a rate less than the

[12] For simplicity, in this example, the assumption elsewhere in this article that d_t is always known at time t has been dropped. It follows that in this example $\delta_t p_t \neq \Delta p_t + d_{t-1} - rp_{t-1}$ but instead $\delta_t p_t = p_t$.

[13] For another illustrative example, consider $\hat{d}_t = \bar{\gamma}\hat{d}_{t-1}+\varepsilon_t$ as with the upper bound for the inequality (11) but where the dividends are announced for the next n years every $1/n$ years. Here, even though $\hat{d}_t$ has the autoregressive structure, ε_t is not the innovation in d_t. As n goes to infinity, $\sigma(\delta p)$ approaches zero.

[14] LeRoy and Porter do assume price as present value of earnings but employ a correction to the price and earnings series which is, under additional theoretical assumptions not employed by Miller and Modigliani, a correction for the double counting.

discount rate and (2) is the solution to the difference equation.[15] With our Standard and Poor data, the growth rate of real price is only about 1.5 percent, while the discount rate is about 4.8%+1.5%=6.3%. At these rates, the value of the firm a few decades hence is so heavily discounted relative to its size that it contributes very little to the value of the stock today; by far the most of the value comes from the intervening dividends. Hence (2) and the implied p^* ought to be useful characterizations of the value of the firm.

The crucial thing to recognize in this context is that once we know the terminal price and intervening dividends, we have specified all that investors care about. It would not make sense to define an *ex post* rational price from a terminal condition on price, using the same formula with earnings in place of dividends.

IV. Time-Varying Real Discount Rates

If we modify the model (2) to allow real discount rates to vary without restriction through time, then the model becomes untestable. We do not observe real discount rates directly. Regardless of the behavior of P_t and D_t, there will always be a discount rate series which makes (2) hold identically. We might ask, though, whether the movements in the real discount rate that would be required aren't larger than we might have expected. Or is it possible that small movements in the current one-period discount rate coupled with new information about such movements in future discount rates could account for high stock price volatility?[16]

The natural extension of (2) to the case of time varying real discount rates is

$$(14)\qquad P_t=E_t\left(\sum_{k=0}^{\infty} D_{t+k}\prod_{j=0}^{k}\frac{1}{1+r_{t+j}}\right)$$

which has the property that $E_t((1+H_t)/(1+r_t))=1$. If we set $1+r_t=(\partial U/\partial C_t)/(\partial U/\partial C_{t+1})$, i.e., to the marginal rate of substitution between present and future consumption where U is the additively separable utility of consumption, then this property is the first-order condition for a maximum of expected utility subject to a stock market budget constraint, and equation (14) is consistent with such expected utility maximization at all times. Note that while r_t is a sort of *ex post* real interest rate not necessarily known until time $t+1$, only the conditional distribution at time t or earlier influences price in the formula (14).

As before, we can rewrite the model in terms of detrended series:

$$(15)\qquad p_t=E_t(p_t^*)$$

where $$p_t^*\equiv\sum_{k=0}^{\infty} d_{t+k}\prod_{j=0}^{k}\frac{1}{1+\bar{r}_{t+j}}$$

$$1+\bar{r}_{t+j}\equiv(1+r_t)/\lambda$$

This model then implies that $\sigma(p_t)\leqslant\sigma(p_t^*)$ as before. Since the model is nonlinear, however, it does not allow us to derive inequalities like (11) or (13). On the other hand, if movements in real interest rates are not too large, then we can use the linearization of p_t^* (i.e., Taylor expansion truncated after the linear term) around $d=E(d)$ and $\bar{r}=E(\bar{r})$; i.e.,

$$(16)\quad \hat{p}_t^*\cong\sum_{k=0}^{\infty}\bar{\gamma}^{k+1}\hat{d}_{t+k}-\frac{E(d)}{E(\bar{r})}\sum_{k=0}^{\infty}\bar{\gamma}^{k+1}\hat{\bar{r}}_{t+k}$$

where $\bar{\gamma}=1/(1+E(\bar{r}))$, and a hat over a variable denotes the variable minus its mean. The first term in the above expression is just the expression for p_t^* in (4) (demeaned). The second term represents the effect on p_t^* of

[15] To understand this point, it helps to consider a traditional continuous time growth model, so instead of (2) we have $P_0=\int_0^{\infty}D_te^{-rt}dt$. In such a model, a firm has a constant earnings stream I. If it pays out all earnings, then $D=I$ and $P_0=\int_0^{\infty}Ie^{-rt}dt=I/r$. If it pays out only s of its earnings, then the firm grows at rate $(1-s)r$, $D_t=sIe^{(1-s)rt}$ which is less than I at $t=0$, but higher than I later on. Then $P_0=\int_0^{\infty}sIe^{(1-s)rt}e^{-rt}dt=\int_0^{\infty}sIe^{-srt}dt=sI/(rs)$. If $s\neq0$ (so that we're not dividing by zero) $P_0=I/r$.

[16] James Pesando has discussed the analogous question: how large must the variance in liquidity premia be in order to justify the volatility of long-term interest rates?

movements in real discount rates. This second term is identical to the expression for p^* in (4) except that d_{t+k} is replaced by $\hat{\bar{r}}_{t+k}$ and the expression is premultiplied by $-E(d)/E(\bar{r})$.

It is possible to offer a simple intuitive interpretation for this linearization. First note that the derivative of $1/(1+\bar{r}_{t+k})$, with respect to $\bar{r}$ evaluated at $E(\bar{r})$ is $-\bar{\gamma}^2$. Thus, a one percentage point increase in $\bar{r}_{t+k}$ causes $1/(1+\bar{r}_{t+k})$ to drop by $\bar{\gamma}^2$ times 1 percent, or slightly less than 1 percent. Note that all terms in (15) dated $t+k$ or higher are premultiplied by $1/(1+\bar{r}_{t+k})$. Thus, if $\bar{r}_{t+k}$ is increased by one percentage point, all else constant, then all of these terms will be reduced by about $\bar{\gamma}^2$ times 1 percent. We can approximate the sum of all these terms as $\bar{\gamma}^{k-1}E(d)/E(\bar{r})$, where $E(d)/E(\bar{r})$ is the value at the beginning of time $t+k$ of a constant dividend stream $E(d)$ discounted by $E(\bar{r})$, and $\bar{\gamma}^{k-1}$ discounts it to the present. So, we see that a one percentage point increase in $\bar{r}_{t+k}$, all else constant, decreases p_t^* by about $\bar{\gamma}^{k+1}E(d)/E(\bar{r})$, which corresponds to the kth term in expression (16). There are two sources of inaccuracy with this linearization. First, the present value of all future dividends starting with time $t+k$ is not exactly $\bar{\gamma}^{k-1}E(d)/E(\bar{r})$. Second, increasing $\bar{r}_{t+k}$ by one percentage point does not cause $1/(1+\bar{r}_{t+k})$ to fall by exactly $\bar{\gamma}^2$ times 1 percent. To some extent, however, these errors in the effects on p_t^* of $\bar{r}_t, \bar{r}_{t+1}, \bar{r}_{t+2}, \ldots$ should average out, and one can use (16) to get an idea of the effects of changes in discount rates.

TABLE 2—SAMPLE STATISTICS FOR PRICE AND DIVIDEND SERIES

	Data Set 1: Standard and Poor's	Data Set 2: Modified Dow Industrial
Sample Period:	1871–1979	1928–1979
1) $E(p)$	145.5	982.6
$E(d)$	6.989	44.76
2) $\bar{r}$	.0480	0.456
$\bar{r}_2$	.0984	.0932
3) $b=ln\lambda$	.0148	.0188
$\hat{\sigma}(b)$	(.0011)	(1.0035)
4) $cor(p, p^*)$	.3918	.1626
$\sigma(d)$	1.481	9.828
Elements of Inequalities:		
Inequality (1)		
5) $\sigma(p)$	50.12	355.9
6) $\sigma(p^*)$	8.968	26.80
Inequality (11)		
7) $\sigma(\Delta p+d_{-1}-\bar{r}p_{-1})$	25.57	242.1
$min(\sigma)$	23.01	209.0
8) $\sigma(d)/\sqrt{\bar{r}_2}$	4.721	32.20
Inequality (13)		
9) $\sigma(\Delta p)$	25.24	239.5
$min(\sigma)$	22.71	206.4
10) $\sigma(d)/\sqrt{2\bar{r}}$	4.777	32.56

Note: In this table, E denotes sample mean, σ denotes standard deviation and $\hat{\sigma}$ denotes standard error. $Min(\sigma)$ is the lower bound on σ computed as a one-sided χ^2 95 percent confidence interval. The symbols p, d, $\bar{r}$, $\bar{r}_2$, b, and p^* are defined in the text. Data sets are described in the Appendix. Inequality (1) in the text asserts that the standard deviation in row 5 should be less than or equal to that in row 6, inequality (11) that σ in row 7 should be less than or equal to that in row 8, and inequality (13) that σ in row 9 should be less than that in row 10.

To give an impression as to the accuracy of the linearization (16), I computed p_t^* for data set 2 in two ways: first using (15) and then using (16), with the same terminal condition p_{1979}^*. In place of the unobserved $\bar{r}_t$ series, I used the actual four–six-month prime commercial paper rate plus a constant to give it the mean $\bar{r}$ of Table 2. The commercial paper rate is a *nominal* interest rate, and thus one would expect its fluctuations represent changes in inflationary expectations as well as real interest rate movements. I chose it nonetheless, rather arbitrarily, as a series which shows much more fluctuation than one would normally expect to see in an expected *real* rate. The commercial paper rate ranges, in this sample, from 0.53 to 9.87 percent. It stayed below 1 percent for over a decade (1935–46) and, at the end of the sample, stayed generally well above 5 percent for over a decade. In spite of this erratic behavior, the correlation coefficient between p^* computed from (15) and p^* computed from (16) was .996, and $\sigma(p_t^*)$ was 250.5 and 268.0 by (15) and (16), respectively. Thus the linearization (16) can be quite accurate. Note also that while these large movements in $\bar{r}_t$ cause p_t^* to move much more than was observed in Figure 2, $\sigma(p^*)$ is still less than half of $\sigma(p)$. This suggests that the variability $\bar{r}_t$ that is needed to save the efficient markets model is much larger yet, as we shall see.

To put a formal lower bound on $\sigma(\bar{r})$ given the variability of Δp, note that (16) makes $\hat{p}_t^*$ the present value of $z_t, z_{t+1}, \ldots$ where $z_t \equiv \hat{d}_t - \hat{\bar{r}}_t E(d)/E(\bar{r})$. We thus know from (13) that $2E(\bar{r})var(\Delta p) \leqslant var(z)$. Moreover, from the definition of z we know that $var(z) \leqslant var(d) + 2\sigma(d)\sigma(\bar{r}\}E(d)/E(\bar{r}) + var(\bar{r})E(d)^2/E(\bar{r})^2$ where the equality holds if d_t and $\bar{r}_t$ are perfectly negatively correlated. Combining these two inequalities and solving for $\sigma(\bar{r})$ one finds

(17)

$$\sigma(\bar{r}) \geqslant \left(\sqrt{2E(\bar{r})}\,\sigma(\Delta p) - \sigma(d)\right)E(\bar{r})/E(d)$$

This inequality puts a lower bound on $\sigma(\bar{r})$ proportional to the discrepancy between the left-hand side and right-hand side of the inequality (13).[17] It will be used to examine the data in the next section.

V. Empirical Evidence

The elements of the inequalities (1), (11), and (13) are displayed for the two data sets (described in the Appendix) in Table 2. In both data sets, the long-run exponential growth path was estimated by regressing $ln(P_t)$ on a constant and time. Then λ in (3) was set equal to e^b where b is the coefficient of time (Table 2). The discount rate $\bar{r}$ used to compute p^* from (4) is estimated as the average d divided by the average p.[18] The terminal value of p^* is taken as average p.

With data set 1, the nominal price and dividend series are the real Standard and Poor's Composite Stock Price Index and the associated dividend series. The earlier observations for this series are due to Alfred Cowles who said that the index is

> intended to represent, ignoring the elements of brokerage charges and taxes, what would have happened to an investor's funds if he had bought, at the beginning of 1871, all stocks quoted on the New York Stock Exchange, allocating his purchases among the individual stocks in proportion to their total monetary value and each month up to 1937 had by the same criterion redistributed his holdings among all quoted stocks. [p. 2]

In updating his series, Standard and Poor later restricted the sample to 500 stocks, but the series continues to be value weighted. The advantage to this series is its comprehensiveness. The disadvantage is that the dividends accruing to the portfolio at one point of time may not correspond to the dividends forecasted by holders of the Standard and Poor's portfolio at an earlier time, due to the change in weighting of the stocks. There is no way to correct this disadvantage without losing comprehensiveness. The original portfolio of 1871 is bound to become a relatively smaller and smaller sample of *U.S.* common stocks as time goes on.

With data set 2, the nominal series are a modified Dow Jones Industrial Average and associated dividend series. With this data set, the advantages and disadvantages of data set 1 are reversed. My modifications in the Dow Jones Industrial Average assure that this series reflects the performance of a single unchanging portfolio. The disadvantage is that the performance of only 30 stocks is recorded.

Table 2 reveals that all inequalities are dramatically violated by the sample statistics for both data sets. The left-hand side of the inequality is always at least five times as great as the right-hand side, and as much as thirteen times as great.

The violation of the inequalities implies that "innovations" in price as we measure them can be forecasted. In fact, if we regress $\delta_{t+1}p_{t+1}$ onto (a constant and) p_t, we get significant results: a coefficient of p_t of $-.1521$ ($t = -3.218$, $R^2 = .0890$) for data set 1 and a coefficient of $-.2421$ ($t = -2.631$, $R^2 = .1238$) for data set 2. These results are

[17] In deriving the inequality (13) it was assumed that d_t was known at time t, so by analogy this inequality would be based on the assumption that r_t is known at time t. However, without this assumption the same inequality could be derived anyway. The maximum contribution of $\bar{r}_t$ to the variance of ΔP occurs when $\bar{r}_t$ is known at time t.

[18] This is not equivalent to the average dividend price ratio, which was slightly higher (.0514 for data set 1, .0484 for data set 2).

not due to the representation of the data as a proportion of the long-run growth path. In fact, if the holding period return H_t is regressed on a constant and the dividend price ratio D_t/P_t, we get results that are only slightly less significant: a coefficient of 3.533 ($t=2.672$, $R^2=.0631$) for data set 1 and a coefficient of 4.491 ($t=1.795$, $R^2=.0617$) for data set 2.

These regression tests, while technically valid, may not be as generally useful for appraising the validity of the model as are the simple volatility comparisons. First, as noted above, the regression tests are not insensitive to data misalignment. Such low R^2 might be the result of dividend or commodity price index data errors. Second, although the model is rejected in these very long samples, the tests may not be powerful if we confined ourselves to shorter samples, for which the data are more accurate, as do most researchers in finance, while volatility comparisons may be much more revealing. To see this, consider a stylized world in which (for the sake of argument) the dividend series d_t is absolutely constant while the price series behaves as in our data set. Since the actual dividend series is fairly smooth, our stylized world is not too remote from our own. If dividends d_t are absolutely constant, however, it should be obvious to the most casual and unsophisticated observer by volatility arguments like those made here that the efficient markets model must be wrong. Price movements cannot reflect new information about dividends if dividends never change. Yet regressions like those run above will have limited power to reject the model. If the alternative hypothesis is, say, that $\hat{p}_t=\rho\hat{p}_{t-1}+\varepsilon_t$, where ρ is close to but less than one, then the power of the test in short samples will be very low. In this stylized world we are testing for the stationarity of the p_t series, for which, as we know, power is low in short samples.[19] For example, if postwar data from, say, 1950–65 were chosen (a period often used in recent financial markets studies) when the stock market was drifting up, then clearly the regression tests will not reject. Even in periods showing a reversal of upward drift the rejection may not be significant.

Using inequality (17), we can compute how big the standard deviation of real discount rates would have to be to possibly account for the discrepancy $\sigma(\Delta p)-\sigma(d)/(2\bar{r})^{1/2}$ between Table 2 results (rows 9 and 10) and the inequality (13). Assuming Table 2 $\bar{r}$ (row 2) equals $E(\bar{r})$ and that sample variances equal population variances, we find that the standard deviation of $\bar{r}_t$ would have to be at least 4.36 percentage points for data set 1 and 7.36 percentage points for data set 2. These are very large numbers. If we take, as a normal range for $\bar{r}_t$ implied by these figures, a ± 2 standard deviation range around the real interest rate $\bar{r}$ given in Table 2, then the real interest rate $\bar{r}_t$ would have to range from -3.91 to 13.52 percent for data set 1 and -8.16 to 17.27 percent for data set 2! And these ranges reflect lowest possible standard deviations which are consistent with the model only if the real rate has the first-order autoregressive structure and perfect negative correlation with dividends!

These estimated standard deviations of *ex ante* real interest rates are roughly consistent with the results of the simple regressions noted above. In a regression of H_t on D_t/P_t and a constant, the standard deviation of the fitted value of H_t is 4.42 and 5.71 percent for data sets 1 and 2, respectively. These large standard deviations are consistent with the low R^2 because the standard deviation of H_t is so much higher (17.60 and 23.00 percent, respectively). The regressions of $\delta_t p_t$ on p_t suggest higher standard deviations of expected real interest rates. The standard deviation of the fitted value divided by the average detrended price is 5.24 and 8.67 percent for data sets 1 and 2, respectively.

VI. Summary and Conclusions

We have seen that measures of stock price volatility over the past century appear to be far too high—five to thirteen times too

[19] If dividends are constant (let us say $d_t=0$) then a test of the model by a regression of $\delta_{t+1}p_{t+1}$ on p_t amounts to a regression of p_{t+1} on p_t with the null hypothesis that the coefficient of p_t is $(1+\bar{r})$. This appears to be an explosive model for which t-statistics are not valid yet our true model, which in effect assumes $\sigma(d)\neq 0$, is nonexplosive.

high—to be attributed to new information about future real dividends if uncertainty about future dividends is measured by the sample standard deviations of real dividends around their long-run exponential growth path. The lower bound of a 95 percent one-sided χ^2 confidence interval for the standard deviation of annual changes in real stock prices is over five times higher than the upper bound allowed by our measure of the observed variability of real dividends. The failure of the efficient markets model is thus so dramatic that it would seem impossible to attribute the failure to such things as data errors, price index problems, or changes in tax laws.

One way of saving the general notion of efficient markets would be to attribute the movements in stock prices to changes in expected real interest rates. Since expected real interest rates are not directly observed, such a theory can not be evaluated statistically unless some other indicator of real rates is found. I have shown, however, that the movements in expected real interest rates that would justify the variability in stock prices are very large—much larger than the movements in nominal interest rates over the sample period.

Another way of saving the general notion of efficient markets is to say that our measure of the uncertainty regarding future dividends—the sample standard deviation of the movements of real dividends around their long-run exponential growth path—understates the true uncertainty about future dividends. Perhaps the market was rightfully fearful of much larger movements than actually materialized. One is led to doubt this, if after a century of observations nothing happened which could remotely justify the stock price movements. The movements in real dividends the market feared must have been many times larger than those observed in the Great Depression of the 1930's, as was noted above. Since the market did not know in advance with certainty the growth path and distribution of dividends that was ultimately observed, however, one cannot be sure that they were wrong to consider possible major events which did not occur. Such an explanation of the volatility of stock prices, however, is "academic," in that it relies fundamentally on unobservables and cannot be evaluated statistically.

Appendix

A. *Data Set 1*: *Standard and Poor Series*

Annual 1871–1979. The price series P_t is Standard and Poor's Monthly Composite Stock Price index for January divided by the Bureau of Labor Statistics wholesale price index (January *WPI* starting in 1900, annual average *WPI* before 1900 scaled to 1.00 in the base year 1979). Standard and Poor's Monthly Composite Stock Price index is a continuation of the Cowles Commission Common Stock index developed by Alfred Cowles and Associates and currently is based on 500 stocks.

The Dividend Series D_t is total dividends for the calendar year accruing to the portfolio represented by the stocks in the index divided by the average wholesale price index for the year (annual average *WPI* scaled to 1.00 in the base year 1979). Starting in 1926 these total dividends are the series "Dividends per share...12 months moving total adjusted to index" from Standard and Poor's statistical service. For 1871 to 1925, total dividends are Cowles series Da-1 multiplied by .1264 to correct for change in base year.

B. *Data Set* 2: *Modified Dow Jones Industrial Average*

Annual 1928–1979. Here P_t and D_t refer to real price and dividends of the portfolio of 30 stocks comprising the sample for the Dow Jones Industrial Average when it was created in 1928. Dow Jones averages before 1928 exist, but the 30 industrials series was begun in that year. The published Dow Jones Industrial Average, however, is not ideal in that stocks are dropped and replaced and in that the weighting given an individual stock is affected by splits. Of the original 30 stocks, only 17 were still included in the Dow Jones Industrial Average at the end of our sample. The published Dow Jones Industrial Average is the simple sum of the price per share of the 30 companies divided by a divisor which changes through time. Thus, if a stock splits two for one, then Dow Jones continues to include only one share but changes the divisor to prevent a sudden drop in the Dow Jones average.

To produce the series used in this paper, the *Capital Changes Reporter* was used to trace changes in the companies from 1928 to 1979. Of the original 30 companies of the Dow Jones Industrial Average, at the end of our sample (1979), 9 had the identical names, 12 had changed only their names, and 9 had been acquired, merged or consolidated. For these latter 9, the price and dividend series are continued as the price and dividend of the shares exchanged by the acquiring corporation. In only one case was a cash payment, along with shares of the acquiring corporation, exchanged for the shares of the acquired corporation. In this case, the price and dividend series were continued as the price and dividend of the shares exchanged by the acquiring corporation. In four cases, preferred shares of the acquiring corporation were among shares exchanged. Common shares of equal value were substituted for these in our series. The number of shares of each firm included in the total is determined by the splits, and effective splits effected by stock dividends and merger. The price series is the value of all these shares on the last trading day of the preceding year, as shown on the Wharton School's Rodney White Center Common Stock tape. The dividend series is the total for the year of dividends and the cash value of other distributions for all these shares. The price and dividend series were deflated using the same wholesale price indexes as in data set 1.

REFERENCES

C. Amsler, "An American Consol: A Reexamination of the Expectations Theory of the Term Structure of Interest Rates," unpublished manuscript, Michigan State Univ. 1980.

S. Basu, "The Investment Performance of Common Stocks in Relation to their Price-Earnings Ratios: A Test of the Efficient Markets Hypothesis," *J. Finance*, June 1977, *32*, 663–82.

G. E. P. Box and G. M. Jenkins, *Time Series Analysis for Forecasting and Control*, San Francisco: Holden-Day 1970.

W. C. Brainard, J. B. Shoven, and L. Weiss, "The Financial Valuation of the Return to Capital," *Brookings Papers,* Washington 1980, *2*, 453–502.

Paul H. Cootner, *The Random Character of Stock Market Prices*, Cambridge: MIT Press 1964.

Alfred Cowles and Associates, *Common Stock Indexes, 1871–1937*, Cowles Commission for Research in Economics, Monograph No. 3, Bloomington: Principia Press 1938.

E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, *25*, 383–420.

______, "The Behavior of Stock Market Prices," *J. Bus., Univ. Chicago*, Jan. 1965, *38*, 34–105.

C. W. J. Granger, "Some Consequences of the Valuation Model when Expectations are Taken to be Optimum Forecasts," *J. Finance*, Mar. 1975, *30*, 135–45.

M. C. Jensen et al., "Symposium on Some Anomalous Evidence Regarding Market Efficiency," *J. Financ. Econ.*, June/Sept. 1978, *6*, 93–330.

S. LeRoy and R. Porter, "The Present Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, forthcoming.

M. H. Miller and F. Modigliani, "Dividend Policy, Growth and the Valuation of Shares," *J. Bus., Univ. Chicago*, Oct. 1961, *34*, 411–33.

F. Modigliani and R. Cohn, "Inflation, Rational Valuation and the Market," *Financ. Anal. J.*, Mar./Apr. 1979, *35*, 24–44.

J. Pesando, "Time Varying Term Premiums and the Volatility of Long-Term Interest Rates," unpublished paper, Univ. Toronto, July 1979.

P. A. Samuelson, "Proof that Properly Discounted Present Values of Assets Vibrate Randomly," in Hiroaki Nagatani and Kate Crowley, eds., *Collected Scientific Papers of Paul A. Samuelson*, Vol. IV, Cambridge: MIT Press 1977.

R. J. Shiller, "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure," *J. Polit. Econ.*, Dec. 1979, *87*, 1190–219.

______ and **J. J. Siegel,** "The Gibson Paradox and Historical Movements in Real Interest Rates," *J. Polit. Econ.*, Oct. 1979, *85*, 891–907.

H. Wold, "On Prediction in Stationary Time Series," *Annals Math. Statist.* 1948, *19*, 558–67.

Commerce Clearing House, *Capital Changes Reporter*, New Jersey 1977.

Dow Jones & Co., *The Dow Jones Averages 1855–1970*, New York: Dow Jones Books 1972.

Standard and Poor's *Security Price Index Record*, New York 1978.

Adoption of Cost-Saving Innovations by a Regulated Firm

By GEORGE SWEENEY*

Regulatory lag is generally credited with providing monetary incentives for the adoption of cost-saving technological changes by regulated firms. Because regulators can not instantaneously adjust price ceilings in response to cost changes, these incentives are inherent to the process of price regulation in a dynamic world. A firm which decreases production costs through technological innovation will enjoy excess profits until the regulators lower price to a level consistent with the new conditions.[1] The longer the delay before regulatory response to a decrease in cost, the greater are the profits which can be derived from a cost reduction, and, therefore, the greater is the incentive for adoption of technological change.[2]

The passive nature of this incentive mechanism is an important characteristic. Although regulatory agencies typically have authority to force price decreases upon firms which do not take advantage of potential cost savings, this power is limited by the difficulty of proving that a firm is laggard, rather than the victim of unfortunate circumstances. Thus, the regulated firm which foregoes potential profit from adopting a cost-saving innovation today may generally reap that profit at some convenient time tomorrow. Indeed, it is the main contention of this paper that, in many circumstances, a regulated monopolist can maximize the present value of profits only by delaying adoption of an innovation. That is, rather than completely adopting a cost-saving innovation when it becomes available, a profit-maximizing regulated firm will choose to adopt the innovation only gradually through time.

The profitability of such delaying procedure may be illustrated by a simple example. Consider a situation in which the price of a firm's product is fixed by a regulatory authority, but periodically adjusted according to the following "cost-plus-markup" scheme. At each review, price for the following period is set equal to average cost of the previous period plus an allowed markup. That is, price in period $i+1$ is set equal to m times average cost in period i ($m \geqslant 1$).

This situation is illustrated in Figure 1. A firm is initially producing output Q_0 at cost $C_0(Q_0)$, with price set by the regulators at $mC_0(Q_0)/Q_0$. Suppose a cost-saving innovation is discovered, which, if employed, would decrease average production cost to $C_1(Q)/Q$. If the firm were to adopt this innovation immediately, it would earn a profit equal to the area of the rectangle $ABCD$ in Figure 1A. At the next regulatory review, price would be lowered to the level m times the new average cost (i.e., to P_1), and profit would fall to that amount represented by the area $AEFG$. This amount of profit would be earned each period thereafter.

However, suppose that the firm were to adopt this innovation in two steps rather than completely adopting the innovation when it first becomes available. This option is sketched in Figure 1B. In the initial period, the firm adopts the innovation throughout approximately half its operations, thereby lowering average cost only to $0H$. In the next period, the firm completes the adoption process, lowering average cost to the final level $0A$. Although this procedure would yield profit equal only to $HICD$ in the initial period, it would result in a price in the next period equal to P_1 (i.e., equal to m times $0H$). Since average cost in that period would fall to $0A$, profit would equal $AJKL$. In the next period, price would be at the new equilibrium level $0G$, and profit would equal $AEFG$. This amount of profit would be earned each period thereafter.

Given the demand and cost functions illustrated in Figure 1, the second alternative

*Assistant professor of economics, Vanderbilt University.

[1] See William Baumol, and Alfred Kahn, ch. 2.

[2] See Elizabeth Bailey.

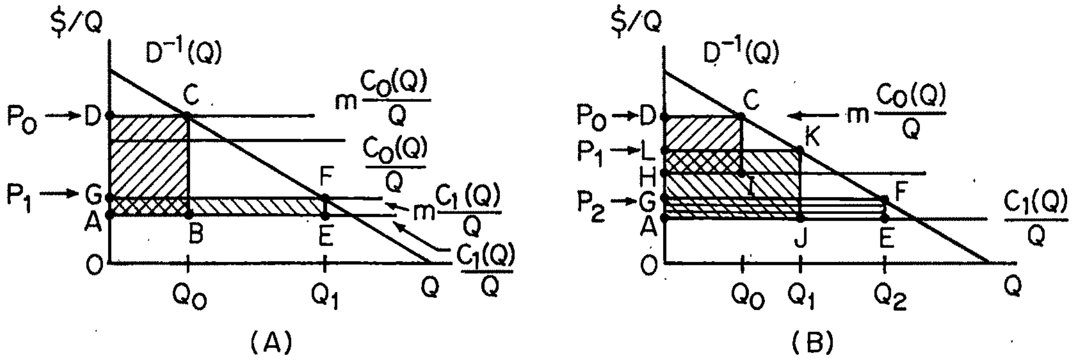


FIGURE 1

yields greater total profit flows. Indeed, greater profits could be earned through even less rapid adoption of the innovation, allowing average cost (and price) to decline by smaller increments in each of a greater number of periods.

Of course, the additional profits from further delay accrue only at a later date. Therefore, the procedure which maximizes the present value of profits depends partly upon the length of time between regulatory reviews and the rate at which the firm discounts future profit flows.

It is the purpose of this paper to illustrate the dynamic aspects of incentives provided by price regulation for adoption of cost-saving innovations. A very simple model of regulation is employed to illuminate characteristics of a regulated firm's profit-maximizing behavior and to determine the importance of parameters of the regulatory environment upon this behavior.

I. Assumptions of the Model

I analyze the profit-maximization problem of a firm subject to cost-plus-markup regulation. Let us assume that the firm produces one homogeneous, nonstorable product. The price of this product is set by a regulatory agency. At the beginning of each period, this price is adjusted to equal the average production cost incurred in the previous period plus a percentage markup. Furthermore, assume that the regulators impose a common carrier obligation. That is, the firm is required to satisfy all demand at the fixed price.

At the beginning of period 0, there becomes available a technological innovation which, if adopted by the firm, would lower production cost at every relevant level of output. Adoption of this innovation would require no adjustment costs or fixed costs. Let us further assume that the firm may choose to take full advantage of this innovation by adopting the innovation throughout its entire operations, or, it may choose to adopt the innovation in any fraction of its operations. It is also assumed that, through partial adoption of the innovation, production cost for any quantity of output can be reduced to any level between the original production cost before the innovation and the production cost with full adoption of the innovation.

Let Q_i represent the rate of production in period i. Let $C(Q)$ designate the minimum potential production cost for a firm which completely adopts the innovation. Let X_i equal the difference between the actual production cost incurred in period i and the minimum potential production cost. Thus, X_i represents the potential cost saving from the innovation, in period i, of which the firm has not taken advantage. By definition, X_i must be nonnegative.

Let P_i represent the price in period i. Assuming the regulatory agency to allow a markup over expenditures of $(m-1)\times 100$ percent, $(m\geqslant 1)$, then the price-setting formula employed by the regulators may be expressed as

$$(1)\qquad P_{i+1}=m\frac{C(Q_i)+X_i}{Q_i}$$

Let us assume that the demand for the firm's output in period i may be represented by the twice continuously differentiable function $D_i(P_i)$. Let $f_i(Q_i)$ represent the inverse-demand function.

And, finally, assume that the firm seeks to maximize the present value of profits derived over the finite, (but arbitrarily large), planning horizon of $N+1$ periods ($i=0, 1,\ldots, N$). Let T_i represent the present value (discounted to the beginning of period 0) of one dollar flow throughout period i. We shall also assume that the rate at which the firm discounts future profits remains constant throughout the planning horizon, and that the periods between regulatory reviews are of equal length.

Letting P_0 represent the (previously determined) price in period zero, the profit-maximization problem of the firm may be expressed mathematically as follows:

Maximize present value of profits:

$$(2)\qquad \sum_{i=0}^{N} T_i[P_iQ_i-C(Q_i)-X_i]$$

subject to

(a) price-setting formula:

$$P_{i+1}=m\frac{C(Q_i)+X_i}{Q_i}\qquad i=0,1,\ldots,N-1$$

(b) common carrier requirement:

$$Q_i-D_i(P_i)=0\qquad i=0,1,\ldots,N$$

(c) nonnegativity constraint:

$$X_i\geqslant 0\qquad i=0,1,\ldots,N$$

(d) initial price condition: $P_i=P_0$ $\quad i=0$

II. Necessary Conditions for Profit Maximization

Determination of the solution to the firm's profit-maximization problem (2) is facilitated by the use of a fictional concept which I term "accounting value." Define accounting value in period, i (a_i), as the product of the markup factor (m) with the expenditure flow in period i.

$$(3)\qquad a_i=m[C(Q_i)+X_i]$$

Because of the nonnegativity of X_i, a_i must be at least as large as $mC(Q_i)$. Except for this minimum bound, the level of accounting value is subject to the choice of the firm: each one dollar increase in expenditure results in an m dollar increase in a_i.

By equation (1) and definition (3), the level of price in each period depends upon the level of accounting value in the previous period. That is,

$$(4)\qquad P_{i+1}=a_i/Q_i$$

Therefore, accounting value is an important determinant of profits. Indeed, the profit-maximization problem of the firm can be viewed as that of choosing an optimal sequence of accounting values. The choice of the optimal level of accounting value in any period may be analyzed with standard marginal techniques. Increasing accounting value in any period requires costs, in the form of excess costs of production, but confers benefits, in the form of an increased price in the ensuing period. If the marginal benefit of accounting value, calculated at the minimum level of accounting value, is smaller than the marginal cost of accounting value, then that minimum level of accounting value is optimal. On the other hand, if the marginal benefit at the minimum level exceeds the marginal cost, then accounting value should be increased up to that level at which its marginal benefit just equals its marginal cost.

In relation to adoption of the innovation, this analysis implies the following. In any period i, the minimum level of accounting value equals $mC(Q_i)$. This level is generated only if the firm completes adoption of the innovation in (or has completed adoption before) period i. Therefore, profit maximization requires that the firm take full advantage of the innovation in period i only if the marginal benefit of accounting value in that period, calculated at $a_i=mC(Q_i)$, is smaller than its marginal cost. If the marginal benefit, at $a_i = mC(Q_i)$, exceeds the marginal cost, the firm should take advantage of the innovation only to the extent that the marginal benefit of accounting value is equated to its marginal cost.

With the help of formulae defining the marginal cost and marginal benefit of accounting value (to be given by equations (5) and (6)), this reasoning suggests an algorithm which may be employed to characterize the profit-maximizing program of innovation adoption. Let us begin by calculating the marginal benefit of accounting value in the initial period, under the assumption of complete adoption of the innovation in that period. If the marginal benefit of accounting value is less then its marginal cost, then complete adoption in the initial period is optimal. Otherwise, this choice is not optimal. If the latter case obtains, we then consider the alternative of only partially adopting the innovation in period zero, and completing the adoption in period 1. Under this scheme, the firm chooses that level of accounting value in period 0 which equates the marginal benefit of accounting value to its marginal cost.[3] Let the resultant price and quantity demanded in period 1 be represented by P_1 and Q_1. Let us calculate the marginal benefit of accounting value in period 1 at the minimum level, $a_1 = mC(Q_1)$. If the marginal benefit is less than or equal to the marginal cost, then this alternative is optimal. Otherwise, this procedure is not optimal, and we next consider the alternative of only partial adoption throughout periods 0 and 1, and completing the adoption in period 2. Under this alternative, levels of accounting value in each of periods 0 and 1 are chosen so as to equate the marginal benefits to marginal costs. Let the resultant price and quantity in period 2 be designated by P_2 and Q_2. The marginal benefit of accounting value in period 2, at the minimum level, $a_2 = mC(Q_2)$, is calculated. If the marginal benefit is less than or equal to the marginal cost, then this procedure is optimal. Otherwise, this procedure can not be optimal, and we must consider the procedure of only partial adoption throughout periods 0, 1, 2, and completion of the adoption in period 3. And so forth.

Implementation of this algorithm requires formulae for the marginal cost and marginal benefit of accounting value. The former may be readily derived. In order to increase accounting value in period i by one dollar, the firm must increase expenditure flow by $1/m$ dollars throughout period i. Therefore, the (present value of) marginal cost of accounting value in period i equals:

$$\$\frac{1}{m}T_i \tag{5}$$

Let H_i represent (present value of) the marginal benefit of accounting value in period i. It is shown in my dissertation that H_i obeys the following recursive relation (primes denote differentiation):

$$H_i = \frac{D'_{i+1}(P_{i+1})}{Q_i}\left\{T_{i+1}\left[P_{i+1} + \frac{Q_{i+1}}{D_{i+1}(P_{i+1})} - C'(Q_{i+1})\right] + \left[mC'(Q_{i+1}) - P_{i+2}\right]H_{i+1}\right\}$$

$$i = 0, 1, \ldots, N-2$$

$$H_{N-1} = \frac{D'_N(P_N)}{Q_{N-1}} \times T_N\left[P_N + \frac{Q_N}{D'_N(P_N)} - C'(Q_N)\right] \tag{6}$$

Let $k+1$ designate the index of the period in which adoption of the innovation is completed[4] ($k \geqslant -1$). The algorithm searches

[3] As indicated by equation (6), the marginal benefit of accounting value in any period depends upon the firm's behavior in later periods. In order to determine the level of accounting value which equates the marginal benefit to the marginal cost, let us assume that no excess expenditure is incurred in any period after the first period in which excess expenditure equals zero. That is, assume that if $X_i = 0$ for $i = k+1$, then $X_i = 0$, for all $i > k+1$. The validity of this assumption must be checked before any procedure is declared "optimal." A method for testing this assumption will be discussed in fn. 6.

[4] Given a finite planning horizon, accounting value in the last period ($i = N$) yields no benefit. Therefore, it will always be optimal to complete adoption of the innovation by period N. That is, k exists, and $k \leqslant N-1$.

over adoption schemes of the following type. In each of the first $k+1$ periods, ($i=0,1,\dots,k$), the marginal benefit of accounting value is equated to its marginal cost. In each of the remaining periods, ($i=k+1,\dots,N$), accounting value is left at its minimum potential level (equal to m times minimum production cost in that period). By equations (5) and (4), this requires that

$$H_i=\frac{1}{m}T_i \qquad i=0,1,\dots,k \tag{7}$$

and

$$P_{i+1}=m\frac{C(Q_i)}{Q_i} \qquad i=k+1,k+2,\dots,N \tag{8a}$$

$$Q_i=D_i(P_i) \tag{8b}$$

For $k>0$, equations (8) may be substituted into equation (6) to yield the following second order difference equation to be satisfied by the sequence of quantities $\{Q_i\}_{i=0}^{k+1}$

$$\frac{1}{Rm}Q_i+D'_{i+1}[f_{i+1}(Q_{i+1})] \times\left[\frac{1}{m}f_{i+2}(Q_{i+2})-MR_{i+1}(Q_{i+1})\right]=0 \qquad i=0,1,\dots,k-1 \qquad (\text{for } k>0) \tag{9}$$

where

$$R\equiv\frac{T_{i+1}}{T_i} \tag{10}$$

$$MR_i(Q_i)\equiv P_i+\frac{Q_i}{D_i'(P_i)} \tag{11}$$

Likewise, for $i=k$, ($k>-1$), equations (8) and (6) imply

$$\frac{Q_k}{Rm}D'_{k+1}[f_{k+1}(Q_{k+1})] \times\left\{MR_{k+1}(Q_{k+1})-C'(Q_{k+1}) +m\left[C'(Q_{k+1})-\frac{C(Q_{k+1})}{Q_{k+1}}\right]\frac{H_{k+1}}{T_{k+1}}\right\}=0 \tag{12}$$

By equation (6), H_{k+1}/T_{k+1} depends upon Q_{k+1} and quantities in periods after $k+1$. But by equations (8), all quantities in periods after $k+1$ depend upon Q_{k+1}. Therefore, equation (12) may be considered to be an equation in only the two variables Q_k and Q_{k+1}. Thus, given a value of k, the sequence of quantities $\{Q_i\}_{i=0}^{k+1}$ may be determined as the solution to the difference equation (9), subject to the terminal condition (12), and the initial condition (13):

$$Q_0=D_0(P_0) \tag{13}$$

The sequence of quantities in the later periods $\{Q_i\}_{i=k+2}^{N}$ is represented by the solution to equations (8).

The solution algorithm is implemented in the following manner. Starting with the initial value equal to -1, a value of k is assumed, and the sequence of quantities which satisfies equations (9), (12), (13), and (8) is calculated.[5] The latter portion of this sequence of quantities is substituted into equations (6) to determine the marginal benefit of accounting value in period $k+1$, when $a_{k+1}=mC(Q_{k+1})$. If the marginal benefit is less than or equal to the marginal cost, (equal to $(1/m)T_{k+1}$), the calculated sequence is optimal.[6] However, if the marginal benefit of accounting value exceeds the marginal cost, the calculated sequence can not be optimal. The trial value of k should be incremented by one, and the procedure repeated.

In practice, this iterative search need not always be performed. In certain cases it is possible to treat the integer k as a variable. The solution to equations (9), (12), (13), and (8) may then be expressed as functions of k, and H_{k+1} calculated as a function of k. The optimal value of k may be chosen as the smallest integer which satisfies the "marginal

[5] For $k=-1$, the sequence of quantities is calculated with only equations (13) and (8). For $k=0$, the sequence is calculated with equations (12), (13), and (8).

[6] In order that this sequence be optimal, it is also necessary that the marginal benefit of accounting value be less than or equal to the marginal cost in each of the periods $i=k+2,k+3,\dots,N$. Therefore H_i must be determined for all $i>k$, and the condition $H_i\leqslant(1/m)T_i$; $i>k$, must be checked.

benefit less than marginal cost" condition:

$$H_{k+1} \leq \frac{1}{m} T_{k+1} \tag{14}$$

This procedure will be illustrated in Section IV.

Having determined the profit-maximizing sequence of outputs, the profit-maximizing sequence of prices may be derived from the inverse demand functions, and the profit-maximizing levels of average production cost may be derived from equations (1). From this information, the optimal cost reduction or amount of innovation adoption in each period may be calculated.

Before illustrating a complete dynamic solution to a firm's profit-maximization problem, I shall analyze the long-run equilibrium behavior of a firm. This analysis will provide a method for determining whether a firm will eventually take full advantage of an innovation.[7]

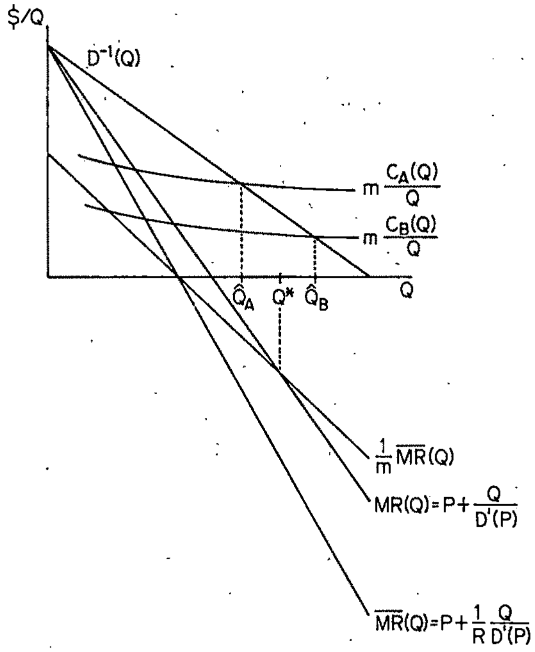


FIGURE 2

III. Long-Run Equilibrium Behavior

I limit the analysis to circumstances in which the demand function does not change during the planning horizon. That is, it shall be assumed that

$$D_i(P_i) = D(P_i) \qquad \text{for all } i$$

I define a long-run equilibrium as a state in which price (and therefore quantity of output demanded and produced) does not change over time.

Suppose that the firm eventually takes full advantage of the innovation. Once the innovation has been completely adopted, price in each period will be determined simply by the demand and average cost functions, according to equations (8). Assuming that the solution to these equations converges, output and price must eventually settle at a point $(\hat{Q}, \hat{P})$ which satisfies the following condition:[8]

$$\hat{P} = m\frac{C(\hat{Q})}{\hat{Q}} \tag{15}$$

where $\qquad \hat{Q} = D(\hat{P})$

The point $(\hat{Q}, \hat{P})$ may be located graphically by the intersection of the demand curve with the locus representing the product of m times the minimum potential average cost. (See Figure 2.)

But if eventual complete adoption of the innovation leads to a long-run equilibrium state at $(\hat{P}, \hat{Q})$, then any strategy of complete adoption can be optimal only if the profit-maximizing conditions are satisfied in the equilibrium state $P_i = \hat{P}$, $Q_i = \hat{Q}$. In this equi-

[7] As stated in fn. 4, profit maximization always requires that the firm completely adopt the innovation by period N. In this section, we search for conditions which ensure that optimal behavior consists of complete adoption before the final period, and that this optimal behavior is independent of the length of the planning horizon.

[8] It is assumed that, in the region of output in which marginal cost exceeds marginal revenue, there exists only one price-quantity pair which satisfies equation (15); define $(\hat{P}, \hat{Q})$ as that price-quantity pair. Also assume that the dynamic system of equations (8) is stable at this point.

librium state, accounting value in each period would equal $mC(\hat{Q})$. By the analysis of Section II, this level of accounting value can be optimal only if the marginal benefit of accounting value at this level is less than or equal to its marginal cost. Therefore, comparison of the steady-state marginal benefit of accounting value with the steady-state marginal cost can indicate whether the firm should eventually take full advantage of the innovation.

The marginal benefit of accounting value in any steady state may be derived by recursively solving equation (6). Letting P^* and Q^* represent the level of price and the quantity of output produced in each period in a steady state, the steady-state marginal benefit of accounting value in period i may be approximated as follows:[9]

$$(16)\quad H_i^* \simeq T_i \frac{MR(Q^*)-C'(Q^*)}{\overline{MR}(Q^*,R)-mC'(Q^*)}$$

where

$$(17)\quad \overline{MR}(Q^*,R) \equiv P^* + \frac{1}{R}\frac{Q^*}{D'(P^*)}$$

In order that the marginal benefit of accounting value be less than the marginal cost at the steady-state equilibrium $(\hat{P}, \hat{Q})$, it is therefore necessary (by equations (5) and (16)) that

$$\frac{MR(\hat{Q})-C'(\hat{Q})}{\overline{MR}(\hat{Q},R)-mC'(\hat{Q})} < \frac{1}{m}$$

Assuming $\hat{Q}$ to be greater than the monopoly quantity, (so that $MR(\hat{Q})<C'(\hat{Q})$), and noting, by definition (17), that $\overline{MR}(Q,R) \leqslant MR(Q)$, we may rewrite the necessary condition for complete adoption of the innovation as follows:

$$(18)\quad \frac{1}{m}\overline{MR}(\hat{Q},R) \leqslant MR(\hat{Q})$$

necessary for eventual complete adoption.

The requirement of condition (18) is represented graphically in Figure 2. Assuming the demand and average cost functions sketched in that diagram, condition (18) is satisfied with the cost function $C_A(Q)$, but is not satisfied with the cost function $C_B(Q)$. That is, the firm would eventually take full advantage of any innovation which would lower production cost to $C_A(Q)$. But the firm would never, (i.e., in no period before N), entirely adopt an innovation which would decrease cost to the lower level $C_B(Q)$.

It is of interest to consider how the regulatory agency's procedures might affect the firm's decision to adopt or not adopt a technological change. The assumptions of the model allow regulatory control of two parameters: the allowed markup determines m, while the frequency of review affects the magnitude of the period-to-period discount factor, R. (Reviewing more frequently reduces the length of time between one period and the next, and therefore increases R.)

The following theorem is proved in the Appendix.

THEOREM 1: *The necessary condition for eventual complete adoption of the innovation* (18) *is satisfied if either of the following conditions hold.*

(a) *The demand function, the minimum cost function, and the allowed markup are such that demand is elastic at price* $\hat{P}$.

(b) *The allowed markup is small enough that* $m<1/R$.

Theorem 1 indicates that the regulatory agency may insure satisfaction of the necessary conditions for complete adoption of an innovation in the long run by either allowing a markup so large that condition (a) is satisfied, or by allowing a markup so small and/or by reviewing so infrequently that condition (b) is satisfied.

[9] This approximation assumes that N is large in comparison to i, and that the following condition is satisfied:

$$\left|\frac{D'(P^*)}{Q^*}[mC'(Q^*)-P^*]\right| < \frac{1}{R}$$

See my dissertation for a more careful discussion.

Of course, whether a firm eventually takes full advantage of the cost saving potential of an innovation should not be the only matter of concern to either regulators or consumers of the firm's product. Costs and prices before a long-run equilibrium is reached must also be considered. In order to measure the effects of changes in regulatory parameters upon disequilibrium behavior, however, it is necessary to determine the complete dynamic solution to the firm's profit-maximization problem.

IV. Example of the Solution Procedure

In this section, I illustrate the use of the solution algorithm described in Section III by solving a simple profit-maximization problem. I hypothesize a firm originally selling Q_0 units of output at price P_0. At date 0, there becomes available a technological innovation which, if adopted, would lower cost at all relevant levels of production. Let us assume the demand function to remain constant throughout the planning horizon, and further assume this demand function to be linear. Also assume minimum potential average cost to be a constant function of output. That is, assume that

$$(19) \qquad D_i(P) \equiv D(P) = Q^A + D'P$$

$$\text{or} \qquad f_i(Q) \equiv f(Q) = -\frac{Q^A}{D'} + \frac{Q}{D'}$$

$$(20) \qquad C'(Q) = \frac{C(Q)}{Q} \equiv C'$$

where D' and C' are constants. Let $\hat{P}$ and $\hat{Q}$ be defined by equations (15). Then equations (20) and (15) imply $\hat{P} = mC'$ and

$$(21) \qquad \frac{\hat{Q}}{D'} - \frac{Q^A}{D'} = mC'$$

Furthermore, equations (19)–(21) imply the following relationships:

$$(22) \qquad C' = \frac{1}{m}\frac{1}{D'}(\hat{Q} - Q^A)$$

$$(23) \qquad MR(Q) = -\frac{Q^A}{D'} + 2\frac{Q}{D'}$$

Equations (19)–(23) may be substituted into equations (9), (12), and (13) to determine conditions for profit-maximizing behavior. These equations imply that, if the firm were to choose accounting value in each of the periods, $i = 0, 1, \ldots, k$, so as to equate the marginal benefit of accounting value to its marginal cost, the sequence of outputs $\{Q_i\}_{i=0}^{k+1}$ would obey the following equations.

The difference equation:

$$(24) \qquad \frac{1}{Rm}Q_i - 2Q_{i+1} + \frac{1}{m}Q_{i+2} + Q^A\left(1 - \frac{1}{m}\right) = 0 \qquad i = 0, 1, \ldots, k-1 \ (\text{for } k \geq 1)$$

the initial condition:

$$(25) \qquad Q_i = Q_0 \qquad \text{for } i = 0$$

and the terminal condition:

$$(26) \qquad \frac{1}{Rm}Q_k - 2Q_{k+1} + Q^A\left(1 - \frac{1}{m}\right) + \frac{1}{m}\hat{Q} = 0$$

It can be shown that equation (26), in conjunction with equation (24), is equivalent to the following condition:[10]

$$(26') \qquad Q_{k+2} = \hat{Q}$$

Furthermore, if accounting value is left at its minimum level (equal to mC') in the remaining periods, $i = k+1, k+2, \ldots, N$, price would be set equal to $\hat{P}$ in period $k+2$ and would remain at that level for the rest of the planning horizon. Therefore, quantity of output would obey the following equation:

$$(27) \qquad Q_i = \hat{Q} \qquad i = k+2, k+3, \ldots, N$$

[10] Proof available upon request, or, see my dissertation.

Employing equations (27) and (21)–(23), the marginal benefit of accounting value in period $k+1$ may be determined from equation (6) as follows:

$$(28)\quad H_{k+1}=T_{k+1}R\frac{1}{Q_{k+1}}\times\left[\hat{Q}\left(2-\frac{1}{m}\right)-Q^A\left(1-\frac{1}{m}\right)\right]$$

According to the analysis of Section II, this behavior can be optimal only if k is the minimum integer for which the marginal benefit of accounting value in period $k+1$ is less than or equal to the marginal cost. By equations (5) and (28), this requires that k be the minimum integer which satisfies the following condition.[11]

$$(29)\quad Q_{k+1}>R\left[\hat{Q}(2m-1)-Q^A(m-1)\right]$$

Thus, the solution algorithm requires that the profit-maximizing sequence of quantities satisfy equations (24), (25), (26′), and (27), and condition (29). The general solution to the difference equation (24) may be represented as follows:

$$(30)\quad Q_i=Q^*+b_1r_1^i+b_2r_2^i$$

where

$$(31)\quad Q^*=\frac{Q^A(m-1)}{2m-1-1/R}=\frac{Q^A(m-1)}{(r_1-1)(1-r_2)}$$

and

$$(32a)\quad r_1=m\left[1+\sqrt{1-1/Rm^2}\right]$$

$$(32b)\quad r_2=m\left[1-\sqrt{1-1/Rm^2}\right]$$

Due to space limitations, we consider only cases in which m exceeds $(1/2)(1+1/R)$, so that both roots r_1 and r_2 are real, and the root r_2 is smaller than one. Analyses of the other possibilities are similar.[12]

The constants b_1 and b_2 in equation (30) may be determined (as functions of k) from the initial condition (25) and the terminal condition (26′). Substituting these values into equation (30), the solution to equations (24)–(26′) may be expressed as follows:

$$(33)\quad Q_i=Q^*+\frac{(\hat{Q}-Q^*)(r_1^i-r_2^i)}{r_1^{k+2}-r_2^{k+2}}+\frac{(Q_0-Q^*)(r_1r_2)^i(r_1^{k+2-i}-r_2^{k+2-i})}{r_1^{k+2}-r_2^{k+2}}$$

Condition (29) may now be employed to determine the optimal value of k. This procedure is facilitated by using definition (31) to rewrite condition (29) as follows:

$$(29')\quad Q_{k+1}>Q^*+(\hat{Q}-Q^*)(2m-1)R$$

Substituting equation (33) with i equal to $k+1$, and noting that (by equations (32)), $r_1r_2=1/R$ and $r_1+r_2=2m$, condition (29′) can be shown to be equivalent to

$$(34)\quad Q_0-Q^*>(\hat{Q}-Q^*)h(k)$$

where

$$(35)\quad h(k)\equiv\frac{R^{k+2}\left[r_1^{k+3}-r_2^{k+3}-\left(r_1^{k+2}-r_2^{k+2}\right)\right]}{r_1-r_2}$$

The implications of condition (34) will be enumerated in two steps. I first consider the case in which Q^* is smaller than $\hat{Q}$, and then consider the opposite case.

Suppose that m and R are large enough that $Q^*<\hat{Q}$. In this case, condition (34) can not be satisfied by any value of k. This may be demonstrated as follows.

[11] If accounting value equals mC' in every period $i>k$, then $P_i=\hat{P}$ and $Q_i=\hat{Q}$ for all $i>k+1$. By equation (6), this implies $H_i/T_i=H_j/T_j$ for all $i,j>k+1$. Therefore, if condition (29) is satisfied, it must also be true that $H_i\leqslant(1/m)T_i$ for all $i\geqslant k+1$. That is, if the marginal benefit of accounting value in period $k+1$ is smaller than its marginal cost, then it will also be true that the marginal benefit of accounting value is smaller than its marginal cost in every period after $k+1$.

[12] Analyses of the remaining cases are available upon request.

For m and R such that $r_2 < 1$, the function $h(k)$ has properties summarized by equation (36).

$$(36) \quad h(k) > 1 \qquad h(k+1) > h(k) \qquad k = -1, 0, 1, \ldots$$

Therefore, the right-hand side of condition (34) is positive, and greater than $(\hat{Q} - Q^*)$ for every k. Furthermore, by assumption, the technological innovation decreases average cost at every quantity of output. Therefore, the pre-innovation quantity Q_0 must be smaller than the post-innovation equilibrium quantity $\hat{Q}$. Thus, $(Q_0 - Q^*) < (\hat{Q} - Q^*)$, and the stated result follows.

The fact that condition (36) cannot be satisfied by any value of k implies that there exists no period in which it is optimal for the firm to complete adoption of the innovation. That is, the firm should never entirely adopt the innovation (except in period N). Alternatively, this fact can be derived from condition (18). Substituting the parameters of the demand and cost functions, condition (18) requires that

$$(37) \quad Q^A(m-1) > \hat{Q}\left(2m - 1 - \frac{1}{R}\right) \quad \text{necessary for eventual complete adoption}$$

But by equations (31), (assuming $r_2 < 1$), this requires that:

$$(38) \quad Q^* > \hat{Q} \quad \text{necessary for complete adoption if } Q^* > 0$$

What is optimal behavior in this situation? Because the planning horizon is finite, the firm should complete adoption of the innovation by the last period. Therefore, the optimal value of k equals $N-1$.[13] The profit-maximizing sequence of quantities is expressed by equation (33), with k equal to $N-1$. Noting that r_2 is less than 1, equation (33) may be approximated as follows:

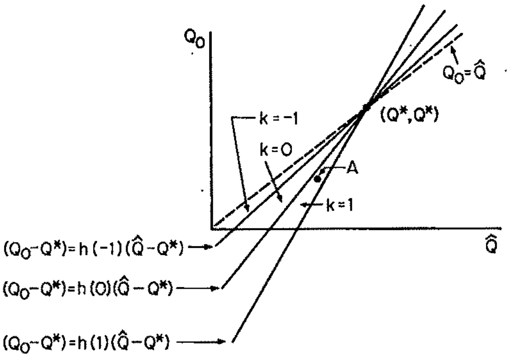

Q_0
$Q_0 = \hat{Q}$
(Q^*, Q^*)
$k = -1$
$k = 0$
A
$k = 1$
$\hat{Q}$
$(Q_0 - Q^*) = h(-1)(\hat{Q} - Q^*)$
$(Q_0 - Q^*) = h(0)(\hat{Q} - Q^*)$
$(Q_0 - Q^*) = h(1)(\hat{Q} - Q^*)$

FIGURE 3

$$(39) \quad Q_i \simeq Q^* + (Q_0 - Q^*) r_2^i \quad \text{for } N \text{ large and } i \ll N$$

That is, assuming the planning horizon to be very long, the optimal sequence of quantities approximately approaches Q^* asymptotically. Note that Q^* represents that quantity at which the marginal benefit of accounting value in the steady state just equals the marginal cost. That is,

$$\frac{MR(Q^*) - C'}{\overline{MR}(Q^*, R) - mC'} = \frac{1}{m}$$

Now suppose that m and R are small enough that Q^* exceeds $\hat{Q}$. The requirements of condition (34) are represented graphically in Figure 3, in which possible values of the initial quantity Q_0 are plotted along the vertical axis. By equations (36), $h(k)$ exceeds 1 for all values of k and increases with k. Therefore, for each value of k, the plot of the equation $(Q_0 - Q^*) = (\hat{Q} - Q^*)h(k)$ is a straight line of slope greater than one passing through the point (Q^*, Q^*). Three such lines are sketched in Figure 3. Given any starting quantity Q_0, and terminal quantity $\hat{Q}$, one may graphically locate the minimal value of k which satisfies condition (34). For example, if the starting and terminal quantities $(Q_0, \hat{Q})$ are represented by the point A in Figure 3, it can be seen immediately that one

[13] From equation (6), it can be seen that the equation expressing the marginal benefit of accounting value in period $k+1$, (equation (28)), is valid only for k such that $k+1 < N$. Therefore, condition (34) applies only for values of k such that $k+1 < N$. As noted in fn. 4, H_N equals zero.

is the minimum value of k which will satisfy the condition that $(Q_0 - Q^*) > h(k)(\hat{Q} - Q^*)$. Therefore, in this case, adoption of the technological innovation would not be completed until period 2 $(i = k+1 = 2)$ and price would reach the equilibrium level $\hat{P}$ only in period 3 $(i = k+2 = 3)$.

As can be seen from examination of Figure 3, the smaller the initial quantity Q_0 in relation to the equilibrium quantity $\hat{Q}$, the greater will be the number of periods before the innovation is completely adopted.

V. Conclusions

The existence of regulatory lag may indeed provide incentives which induce regulated firms to eventually adopt cost-saving technological changes. However, I have shown, by example, that these same incentives may induce the firm to retard the speed at which it adopts such innovations. Furthermore, I have shown that there exist circumstances under which regulatory lag can not provide sufficient incentive to induce eventual complete adoption of an innovation.

The nature of the adoption scheme which maximizes the present value of a regulated firm's profits depends critically upon the characteristics of the regulatory procedure, including the tightness of the regulatory constraint (the size of the markup), and the magnitude of the regulatory lag (the frequency of regulatory review). Optimal behavior also depends upon the nature of the demand function and its changes over time. Thus, one may conclude that the incentives for adopting innovations provided by regulatory lag may be far more complex than casual discussions of the topic suggest.

Appendix: Proof of Theorem 1

(a) Noting that $1/R \geq 1$ and $Q/D' < 0$, definitions (11) and (17) imply that $\overline{MR}(Q, R) \leq MR(Q)$. Therefore:

$$\text{(A1)} \qquad \frac{1}{m}\overline{MR}(Q, R) \leq \frac{1}{m} MR(Q)$$

By assumption, $1/m \leq 1$. Therefore:

$$\text{(A2)} \qquad \frac{1}{m} MR(\hat{Q}) \leq MR(\hat{Q})$$

$$\text{for } \hat{Q} \text{ such that } MR(\hat{Q}) > 0$$

Equations (A1) and (A2) together imply

$$\frac{1}{m}\overline{MR}(\hat{Q}, R) \leq MR(\hat{Q})$$

$$\text{for } \hat{Q} \text{ such that } MR(\hat{Q}) > 0$$

(b) Substituting the definitions of $MR(Q)$ and $\overline{MR}(Q, R)$, condition (18) requires

$$\text{(A3)} \qquad \frac{Q}{D'(P)}\left(\frac{1}{m}\frac{1}{R} - 1\right) \leq P\left(1 - \frac{1}{m}\right)$$

The right-hand side of condition (A3) is nonnegative. Therefore, this condition is satisfied for any m and R such that $((1/m)(1/R) - 1) > 0$. That is, inequality (18) is satisfied if $1/R > m$.

REFERENCES

W. J. Baumol, "Reasonable Rules for Rate Regulation: Plausible Policies for an Imperfect World," in Paul W. MacAvoy; ed., *The Crisis of the Regulatory Commissions*, New York 1970.

E. E. Bailey "Innovation and Regulation," *J. Public Econ.*, Aug. 1974, *3*, 285–95.

Alfred Kahn, *The Economics of Regulation*, Vol. 2, New York 1971.

G. H. Sweeney, "A Dynamic Theory of A Firm Subject to Regulation," unpublished doctoral dissertation, Northwestern Univ. 1978.

Price Controls and the Behavior of Auction Markets: An Experimental Examination

By R. Mark Isaac and Charles R. Plott*

Price ceilings and price floors are common in all market systems. The ancient Greeks and Hellenistic era Egyptians are known to have utilized price controls (see H. Michele, p. 272, and J. P. Levy, p. 41), and numerous public policy questions today involve them. Apparently for as long as price controls have existed, their effects have been debated. For example, Diocletion's favorable view of his price ceilings[1] was disputed by the religious philosopher, Lactantius, who charged that the policy led to "scarcity and...low grade articles" (p. 145).

The standard partial-equilibrium theory about the effects of price controls, the theory which is subjected to so much criticism, does not seem to have changed since Leon Walras. It is applied widely to a variety of market institutional arrangements including auction markets such as those studied below. If the demand schedule is downward sloping and if the supply schedule is increasing as shown in Figure 1, there should be an equilibrium price-quantity pair of ($.60,20). Nonbinding price controls, such as a price ceiling at or above the equilibrium or a price floor below equilibrium, should have no effects on the market. If the controls are binding, such as a price ceiling at $.55 or a price floor at $.70, then the market achieves an inefficient price-quantity pair with the market price equaling the controlled price.

However, in spite of its prominent textbook status, the applicability of the model is questioned regularly.[2] Criticisms range from complete rejections of economics to elaborate theories of collusion. As an example of the latter, consider the "focal point" hypothesis as found in F. M. Scherer (p. 352). Perhaps the price ceiling will act as a focal point. Sellers, by focusing on a nonbinding ceiling, may be able to tacitly collude to keep prices above the equilibrium. Thus, the otherwise nonbinding price ceiling can have an effect on prices. A similar theory can be advanced about the effects of price floors. For us the existence of this general controversy and the focal point hypothesis regarding the dynamic effects of price controls seemed sufficient to justify a systematic examination of the subject.

The objectives of this study are to examine the applicability and/or accuracy of the textbook model as applied to laboratory auction markets. Our hope is that by studying the implications of price controls in simple controlled settings, we will be in a better position to analyze more complicated markets which have been the traditional subjects of academic and scientific concern. The choice to study auction markets, as opposed to other forms of market institutions, reflected an attempt to maintain continuity with other experimental studies. Our results are not exactly what we expected and they probably raise more questions than they answer.

*University of Arizona and California Institute of Technology, respectively. The financial support of the National Science Foundation and the Caltech Program for Enterprise and Public Policy is gratefully acknowledged. We thank Julie Laherty for her comments on an earlier draft.

[1] The following quotation is excerpted from *Roman Civilization* (pp. 464–66):

> In response to the needs of mankind itself, which appears to be praying for release, we have decided that maximum prices of articles for sale must be established. We have not set down fixed prices, for we do not deem it just to do this, since many provinces occasionally enjoy the fortune of welcome low prices....
>
> It is our pleasure, therefore, that the prices listed in the subjoined schedule be held in observance in the whole of our Empire. And every person shall take note that the liberty to exceed them at will has been ended, but that the blessing of low prices has in no way been impaired in those places where supplies actually abound....
>
> Emperor Diocletian, *The Edict on Prices*, A.D. 301

[2] During the course of preparing this paper, we noted several heated local political discussions concerning "fair trading" of liquor products, rent ceilings, and wage floors for municipal employees.

TABLE 1

		Series I		Series II		Mixed	
		No Experience		Experience		Experience	
		Experiment	Period	Experiment	Period	Experiment	Period
No Controls		I	all	III	1–3, 7	II	all
		VII	9	VIII	9		
		IX	9–10	X	9–10		
		XII	9–11	VI	7–11		
Controls at	Price ceiling	IV	all				
Equilibrium	at equilibrium	V	all				
	Price floor						
	at equilibrium			VI	1–6		
Nonbinding	Price ceiling 5¢			III	4		
Controls	above equilibrium	VII	1–8	VIII	1–8		
	Price floor 5¢						
	below equilibrium	IX	1–8	X	1–8		
	Price ceiling 10¢						
	above equilibrium	XI	all				
Binding	Price ceiling 10¢			III	5–6		
Controls	below equilibrium	XII	1–8	VIII	10		

I. Experimental Design

A total of twelve experimental sessions were conducted. These are listed in Table 1 according to the subject's laboratory market experience and according to the price-control institution imposed. The instructions were those of Plott and Vernon Smith (Appendix, pp. 147–52) and Ross Miller, Plott, and Smith (Appendices, pp. 610–21) with a price ceiling (floor) provision added as indicated below. Participants in Series I (recruited from Pasadena City College) had no previous experience in laboratory markets. All participants in Series II (recruited from Caltech) had participated in at least one other laboratory market with parameters differing from the experiments reported here.

The laboratory design of each experimental session consisted of an auction market with four buyers and four sellers. Preferences were induced following the theory of induced preference (see Smith; Plott). Buyers made money by buying from the sellers and reselling to the experimenter according to prespecified terms. Likewise, sellers made money by buying from the experimenter at prespecified costs and reselling to the buyers. In addition, each individual received a five-cent trading commission. The value of the redemption values for each individual is indicated on Figure 1.

Each market involved a series of "trading periods" in which market participants were free to buy and sell. The individual parameters were identical each period. By application of the theory of induced preference (and/or derived demand) the individual parameters become limit prices which can be "summed" in accord with competitive market theory to produce the demand and supply curves represented in Figure 1. These curves remained constant over all periods and, except for small shifts upward by a constant, indicated below, were the same across all experiments.

Markets were organized as two-sided oral auction markets. All participants had free access to the market floor to make bids to buy (offers to sell) or to accept any outstanding offer (bid). Each bid canceled previous bids, and offers canceled previous offers. All ties were broken by random process.

The institutions being examined are a series of price ceilings and price floors. Specifically, the following paragraph is an example of a price ceiling: "During this experiment, no bids or offers may be made or accepted at a

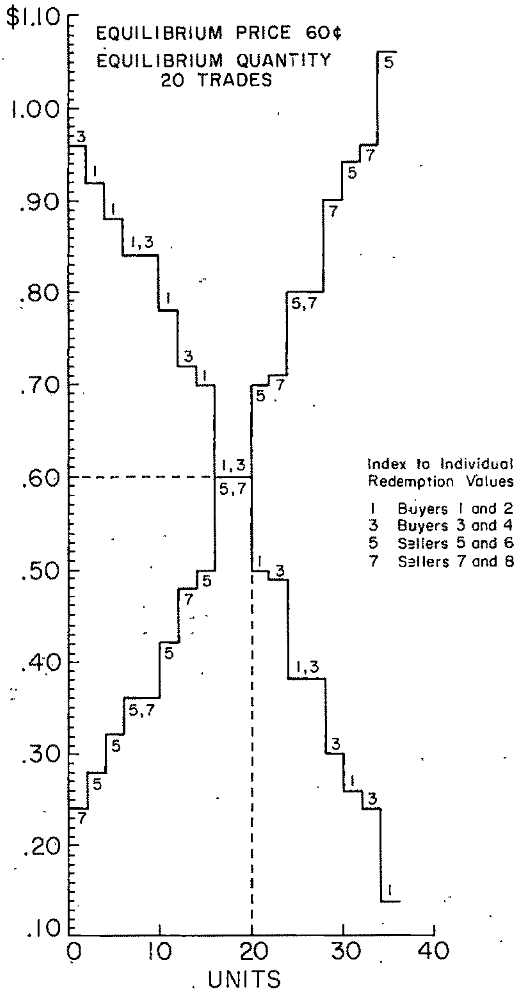


FIGURE 1

price greater than__cents. Of course, you may still make or accept bids or offers at a price less than or equal to this amount."

In general, our experiments can be divided into seven categories as follows ($\bar{P}=$ maximum price, $\underline{P}=$minimum price, $P_0=$ competitive equilibrium):

(1) no price controls
(2) & (3) price controls precisely at predicted equilibrium ($\bar{P}=P_0$; $\underline{P}=P_0$)
(4) & (5) strictly nonbinding price controls ($\bar{P}>P_0$; $\underline{P}<P_0$)
(6) & (7) strictly binding price controls ($\bar{P}<P_0$; $\underline{P}>P_0$).

Not all categories were examined because of the expense and the nature of the evidence obtained from the experiments we did run (see Table 1).

The results of the twelve experimental sessions are presented in the following section, with a particular emphasis upon the patterns which exhibit regularity, and upon the relationship between these results and the existing theoretical literature. Additionally, we will consider the significance of our results for future research.

We have focused the study on the following three aspects of market behavior:

1) Price Levels and Market Volume: Price level refers to the average price of a contract during a period. Sometimes the range of prices during a period is referenced. Volume refers to the number of contracts during a period.

2) Market Responses to Institutional Modifications: During the course of several experiments price controls were removed. Occasionally a control was added or changed (see Table 1).

3) Efficiency: The efficiency index developed by Plott and Smith is used here. Markets are 100 percent efficient if and only if the total of subjects' profits and commissions is maximized during a trading period. The efficiency is the actual sum of subjects' profits and commissions divided by the theoretical maximum of this sum. This measure is related to the maximum of consumer's plus producer's surplus.

II. Experimental Results: Some Preliminary Conclusions

We can report two major results and a conjecture. The results are: First, that market behavior under price controls is more closely approximated by the competitive model than by the focal point model; and secondly, that markets under price controls exhibit behavioral regularities which are not included in standard analyses and some of which cannot be explained by the "traditional" competitive model. Specifically, four such regularities were noted: (i) controls at the competitive equilibrium cause market prices to diverge from the competitive equilibrium; (ii) the removal of nonbinding controls induces

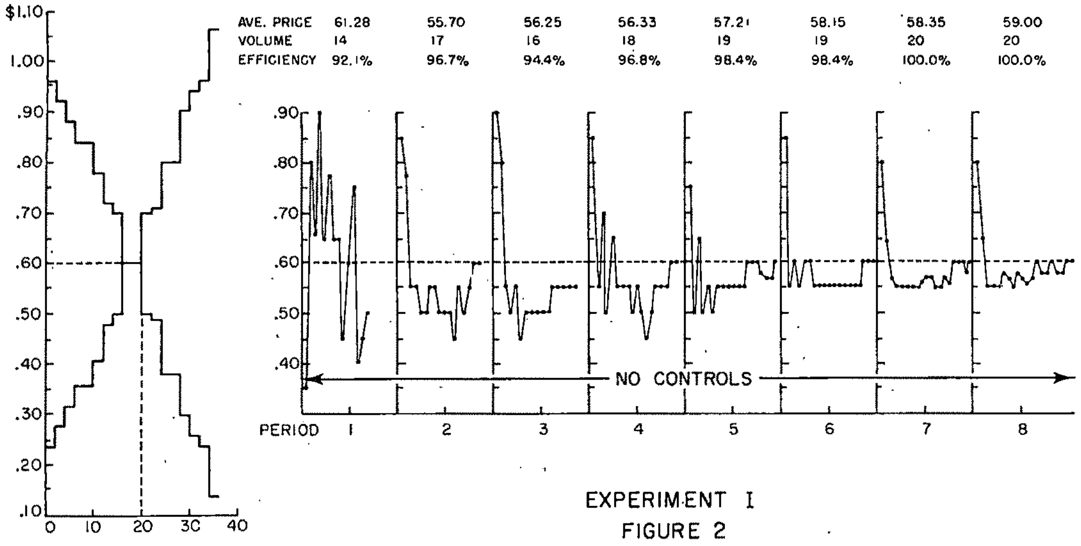


FIGURE 2

changes in market prices; (iii) inefficiencies induced by binding controls are greater than those predicted by the standard application of consumer's surplus analysis. The amount of additional loss depends upon the method of resolving the rationing problem; and finally, (iv) adjustment of prices when binding controls are removed appears to involve an initial discontinuity or "jump" rather than a continuous adjustment. The conjecture is that nonbinding controls act like a "buffer" which holds prices below (above) the "natural" market equilibrium in the case of price ceilings (floors).

Since the two results can be easily demonstrated, we have organized the following subsections, which contain a more detailed examination of the data, in a manner which highlights the nature of the conjecture. First, we discuss the behavior of markets with no controls at all. It is here that the concept of a natural equilibrium (as opposed to the equilibrium point of the competitive model) is explored. The second and third subsections, respectively, address the results of experiments with nonbinding controls and binding controls.

The experimental results are displayed in Figures 2 through 13. Shown in these figures are all contract prices arrayed according to the order (in time) in which the contract occurred. The dotted line always indicates the competitive model equilibrium price (in the absence of controls). During some experiments institutional changes were made, for example, a price control may have been removed or imposed. A double line separates the periods where institutional changes are initially imposed and the nature of the change is indicated on the figure. The equilibrium price, average prices, volume, and efficiencies for each period are on the figures.

A. *No Price Controls*

Three experiments were conducted with no price controls at all. These are Experiments I, II, and III (periods 1, 2, 3, and 7) on Figures 2, 3, and 4. In addition, price controls were removed for selected periods in other experiments (see Table 1).

Laboratory markets (including those examined here), when organized as a "double oral auction"[3] without price controls, invariably exhibit the following properties. These properties are important since they serve as standards against which the effects of price controls can be judged. (a) Efficiencies are high and approach 100 percent and stabilize

[3] We refer specifically to those in which, as here, small trading commissions are paid.

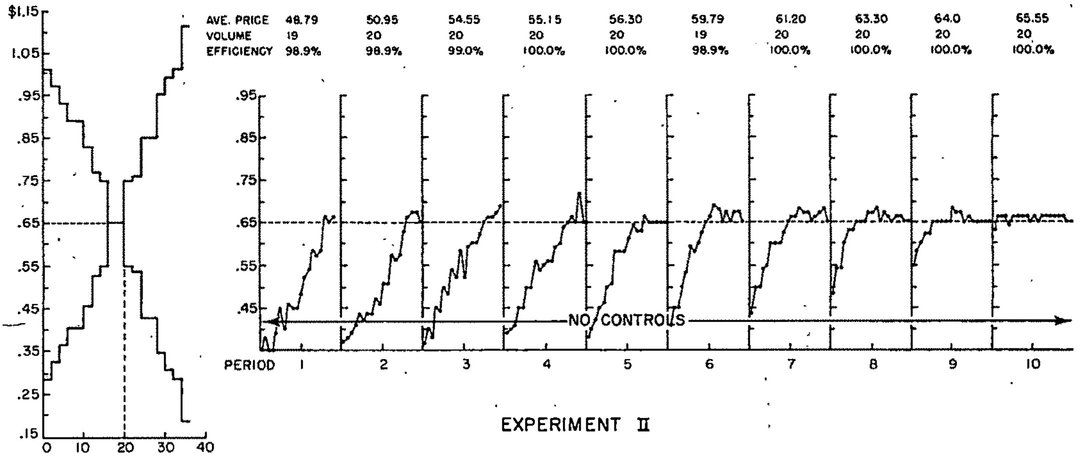


FIGURE 3

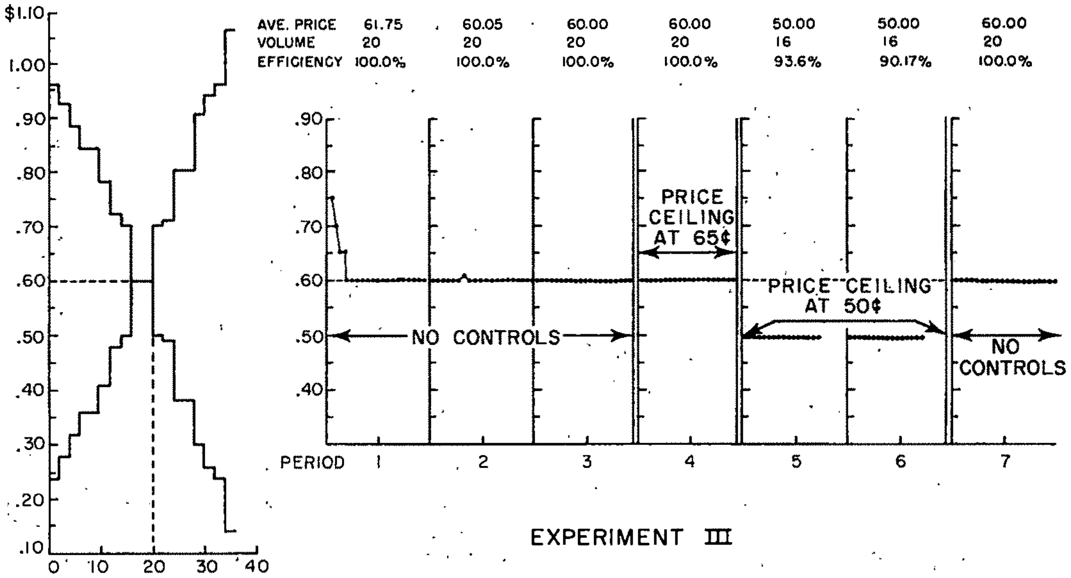


FIGURE 4

once high efficiencies are achieved (i.e., above 98 percent). (b) The variance of prices tends to diminish with replications of periods. (c) If there are many trades at prices other than the equilibrium, they tend to be on both sides of the equilibrium. (d) Average prices tend to stabilize near the competitive equilibrium price.

Experiment III (Figure 4) dramatically demonstrates the frequently observed power of the competitive model. Prices converge almost immediately to the competitive price with zero variance and 100 percent efficiency. While subjects in this experiment did not know the market parameters, they had all had previous experience in laboratory markets. Subject experience is suspected to be a primary reason for the relatively rapid convergence and low variance of Experiment III relative to the other two no-control experiments (I and II).

Sometimes markets have sellers (buyers) who are willing to sell (buy) units at prices considerably below (above) the equilibrium price even though many trades occur at or above (below) equilibrium. These individuals, who do not seem inclined to "hold out" for one of the better deals are called "relatively soft" sellers (buyers). In Experiment I (Figure 2) notice that the first trade or two in every period is considerably above the other trades. All of these contracts involved the same "soft" buyer. In Experiment II (Figure 3) notice that many low-priced contracts occur at the beginning of each period. These all involved the same two soft sellers whose anxiousness to sell resulted in low contract prices. Exactly why this occurs is not known (in Experiment II, however, one of the soft sellers had no previous experience in laboratory markets) but whatever the reason the behavior is usually "corrected" by the last few periods. It is important to notice, however, that "softness" seems to affect neither the market efficiency (in all three experiments it is over 98 percent by the fifth period and increasing) nor the tendency for trades to occur on both sides of equilibrium. Properties (a), (b), and (c) are exhibited in all three experiments. However, to the extent that the average prices diverge from the equilibrium of the competitive model, we need a concept of a natural equilibrium. The effects of price controls then must be gauged relative to this natural tendency as opposed to the prediction of the competitive model.

The major difficulty with supporting our buffer conjecture above can now be made clear. Indeed the soft trader problem is the reason the result is listed as a conjecture instead of a conclusion. If soft buyers or sellers exist, the average price may remain removed from the competitive equilibrium price. Thus the influence of price controls must be measured against this natural tendency rather than the equilibrium of the model. But the natural tendency cannot be known until the market operates and since the softness of subjects may be modified by any market experience, the very act of observing the "natural equilibrium" which differs from that of the competitive model may cause it to change. Thus, there is currently no "fixed" measure against which the influence of price controls can be identified.

Our initial experimental design was not constructed to deal with this difficulty. At best we are able to establish within our design the plausibility of the buffer conjecture and identify certain properties of the buffer phenomenon if indeed it exists.

B. *Nonbinding Controls*

Nonbinding price controls existed in all or parts of eight of the twelve experimental markets. The first experiments, reported here as Experiment IV (Figure 5) and Experiment V (Figure 6), involved a price ceiling at the competitive equilibrium price and Experiment VI (Figure 7) involved a floor at the competitive equilibrium. The results from these three experiments led to additional experiments with nonbinding controls "near" the equilibrium price (Experiments VII–XI on Figures 8–12, respectively). These will be covered in order below.

Two conclusions can be supported by a reference to all experiments with nonbinding controls. First, the market behavior under nonbinding price controls is more closely approximated by the competitive model than the focal point model advanced in the introduction. In *no* period of any experiment is the average market price closer to the price control than the competitive equilibrium price. When the ceiling is equal to the competitive equilibrium price, the average prices tend to diverge from the ceiling. When the nonbinding price control is not equal to the competitive equilibrium price, the average price (indeed the entire range of prices) of every period is closer to the competitive equilibrium. The rejection of the focal point model in favor of the competitive equilibrium model seems amply justified.

The second conclusion, on the other hand, highlights a possible incompleteness in the traditional model. Removal of a nonbinding price control affects the price level. The action seems to "disequilibrate" the market. Nonbinding price controls are removed in Experiments VI–X (Figures 7–11, respectively). In every case the removal of the nonbinding control is followed by a movement in the average price. The only case where the spirit of this conclusion is violated is Experiment III, period 4 (Figure 4) in which the nonbinding control was added after

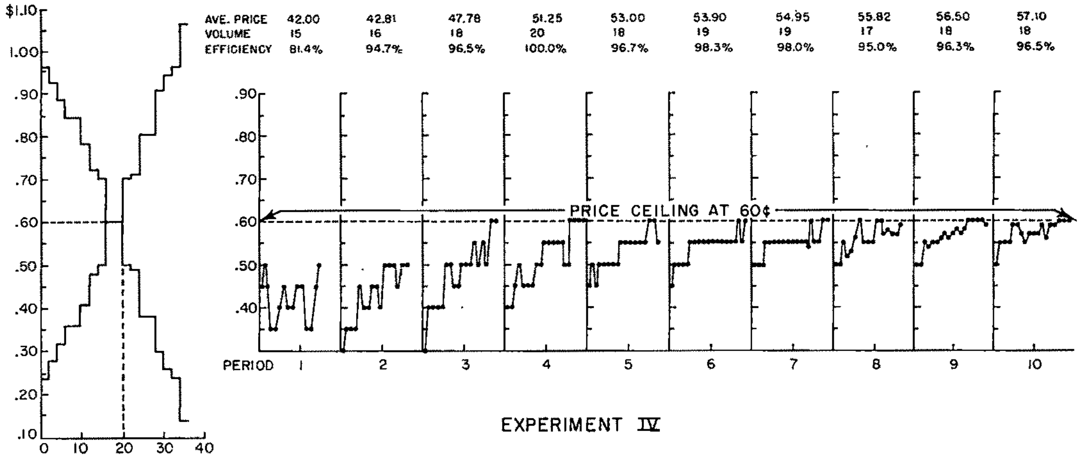


FIGURE 5

the market had already converged and induced no changes at all in the level of prices.

According to traditional models the equilibrating properties of markets depend only upon the magnitude of excess demand. Since the removal of nonbinding price controls does not affect the magnitude of excess demand, the traditional model cannot account for the resulting changes in the price level. Exactly how the traditional model must be supplemented is not clear. Perhaps the removal of controls makes available additional strategies to one side or the other, thereby giving differential advantages. Perhaps any "announcement" in experimental markets will cause "disequilibrations." Perhaps the change creates additional uncertainty, thereby encouraging additional search activity by some participants and conservative or soft trading on the part of the others. Clearly, both additional theory and experiments are needed before the reasons for the phenomenon can be identified.

We turn now to the conjecture, the "buffer hypothesis" by examining first the experiments with price controls placed at the competitive equilibrium (Experiments IV–VI on Figures 5–7). In the price ceiling experiments, IV and V, prices are almost stabilized at an average below the ceiling with few trades at the competitive equilibrium ceilings. Efficiencies remain below 98 percent with marginal units not being traded even though in Experiment IV an efficiency of 100 percent was attained once during an early period. In the price floor experiment, VI (with experienced subjects), prices converged to the floor in a manner seemingly contradictory to the buffer hypothesis, but when the floor was removed (period 7), prices immediately dropped to a lower level. Thus, in the context of the buffer hypothesis the natural equilibrium was below the competitive equilibrium for this group of subjects. Efficiencies in this experiment approximate 100 percent.

Four experiments (VII–X) were conducted with nonbinding controls placed within five cents of the competitive equilibrium. The buffer hypothesis can be applied to all four sessions. The evidence is strongest for Experiments VII–IX where trades seldom if ever occur at prices between the price control and the competitive equilibrium. When the control is removed, prices immediately rise (fall) to above (below) the competitive equilibrium in the case of price ceilings (floor). In Experiments VII and IX the efficiency level did not behave in the stable manner characteristic of markets without controls. Instead, the efficiency sometimes attained the 98 percent level but did not remain. Experiment X differs because prices converged initially below the competitive equilibrium but even in this experiment prices fell when the nonbinding

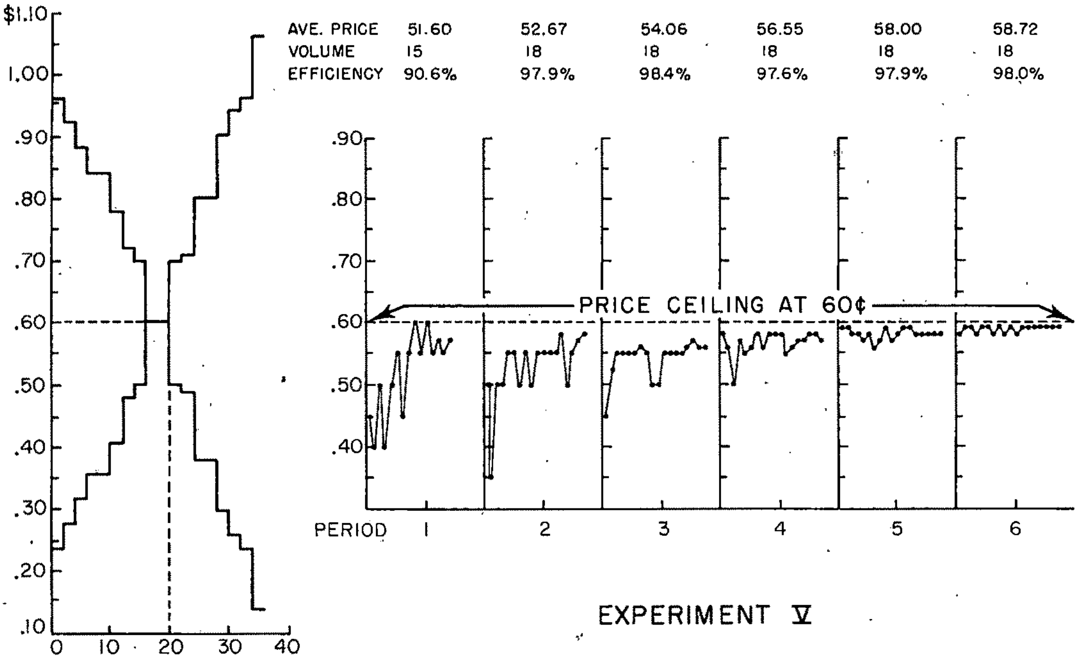


FIGURE 6

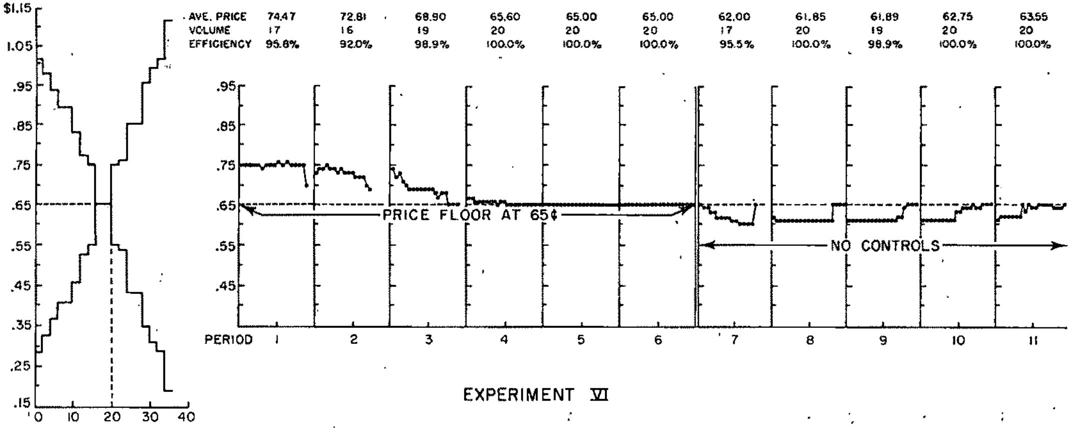


FIGURE 7

floor was removed. Thus, for this experiment application of the buffer hypothesis must assume that the sellers were soft and the nonbinding floor acted to hold prices above the natural equilibrium.

In Experiment XI (Figure 12) a nonbinding ceiling was placed ten cents above the equilibrium. Since prices here converged very close to the competitive equilibrium and since the control remained throughout the whole experiment, we have little to say about it. We suspect, however, that the buffer effect is weak at best here where the control is "far" from the equilibrium price.

As indicated above, we can at best speculate about the reasons for the buffer effect. It may have something to do with the information and "search." The results of Experiment

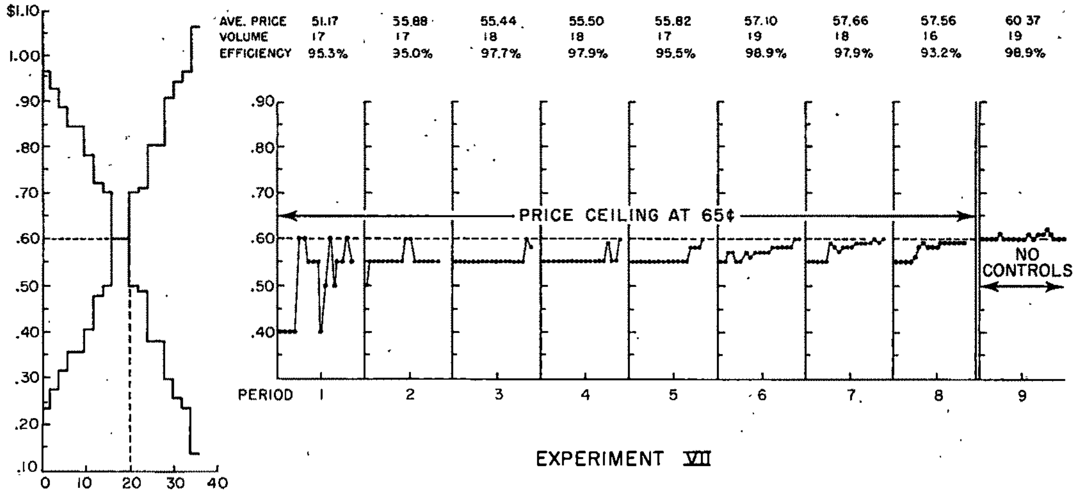


FIGURE 8

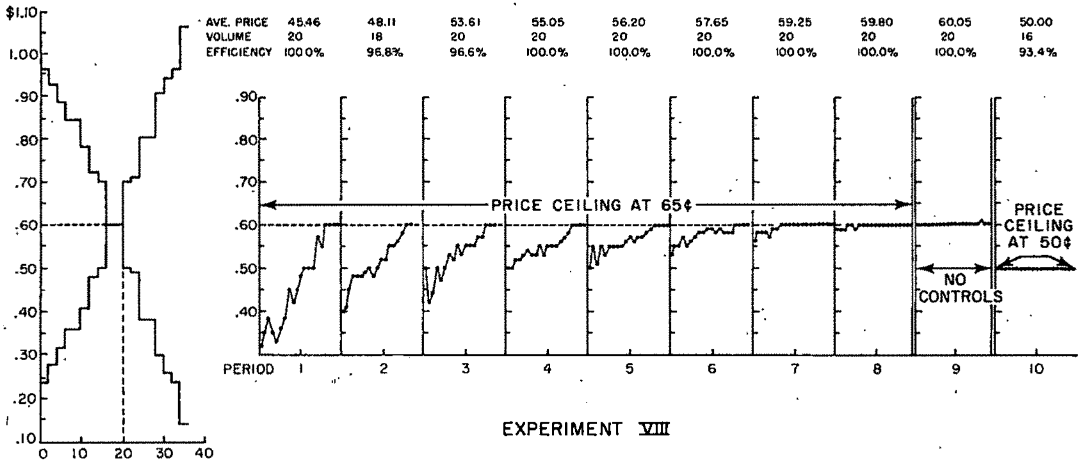


FIGURE 9

III, period 4, were revealing in this respect. Adding a nonbinding ceiling there made no difference at all.

C. *Binding Price Controls*

For the first eight periods of Experiment XII (Figure 13) a price ceiling of fifty cents existed which was below the competitive equilibrium price of sixty cents. The ceiling was removed after the eighth period. In period 10 of Experiment VIII (Figure 9) and in periods 5 and 6 of Experiment III (Figure 4) a price ceiling below the equilibrium was imposed.

The experiments were motivated by the buffer hypothesis. Perhaps the buffer would work to keep prices below a *binding* control. In this respect the control could be viewed as the opposite of the focal point hypothesis as introduced above. Perhaps the ceiling (floor) acts as a signal to the buyers (sellers) and helps them coordinate to hold prices below (above) the ceiling (floor).

As can be seen from all the figures, this alternative hypothesis seems to be wrong.

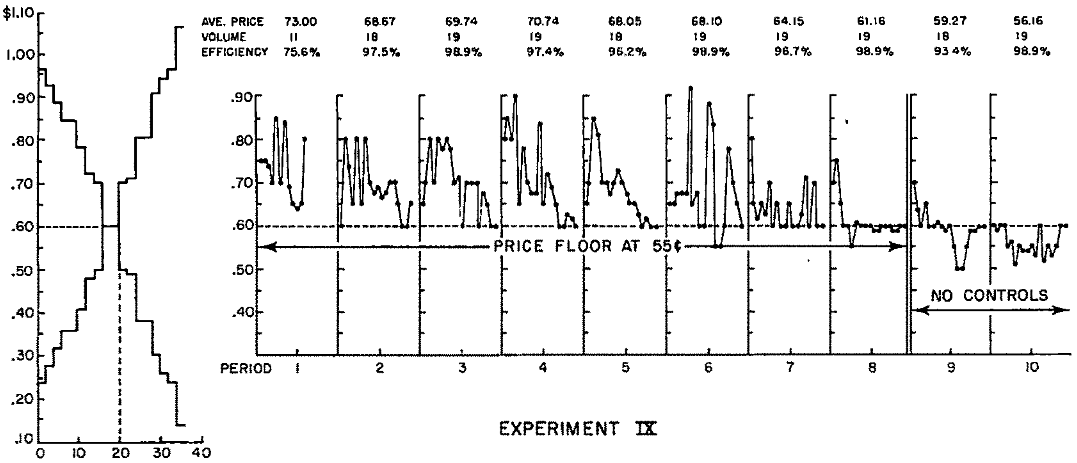


FIGURE 10

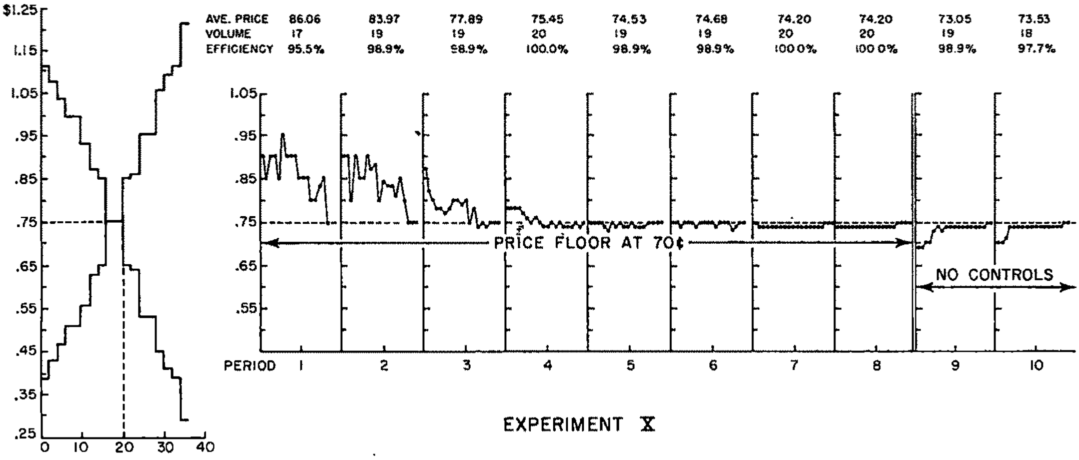


FIGURE 11

Prices converge rapidly to the binding ceilings. Price equals the ceiling and the volume equals the competitive supply function evaluated at the price ceiling. For the case of binding controls the market price behaves as predicted by the traditional model.

In the course of these experiments we discovered two modes of behavior we did not anticipate. The first "unexpected" results occurred when the controls were removed. The adjustment *path* of prices when the binding ceiling is removed differs somewhat from the standard dynamic hypothesis. In period 9 of Experiment XII the binding price ceiling was removed. The mean price jumped immediately to more than thirteen cents *above* equilibrium and then converged down toward equilibrium rather than adjusting continuously upward as suggested by most dynamics models. A discontinuity in adjustment was also present when a binding price ceiling was added in Experiment III (periods 5 and 6) and then removed (period 7). In this market (in which subjects were experienced) prices simply adjusted back immediately to the previously attained equilibrium without "overshooting." This latter result suggests that information, in addition to possibly the

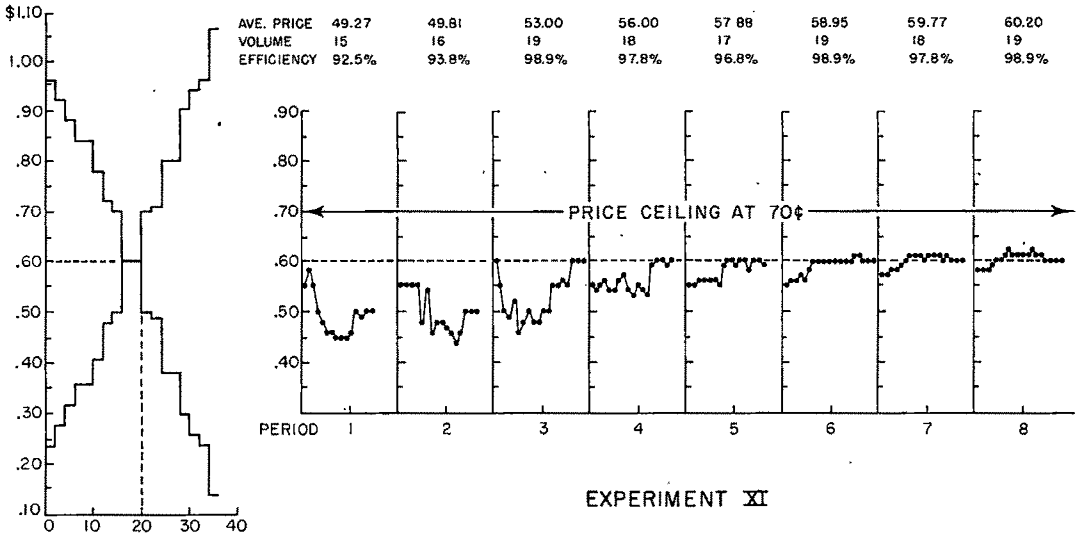


FIGURE 12

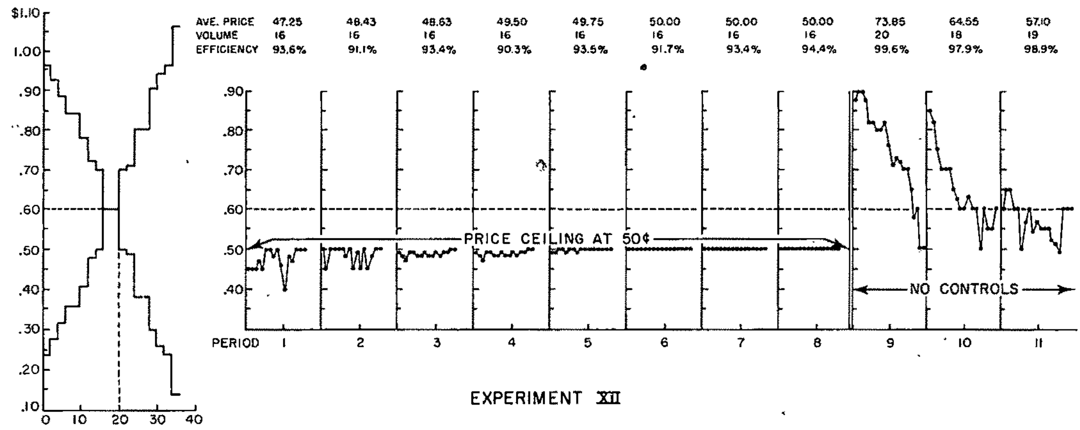


FIGURE 13

magnitude of excess market demand, plays a systematic role in the formation of adjustment paths. Of course more experimentation and theory are necessary.

Secondly, analysis of these experiments with binding controls reveals a source of inefficiency not often stressed in the economics literature. Efficiency losses can result from both the price ceiling as well as the choice of the rationing process used in conjunction with binding price controls. Because of the fifty-cent price ceiling in Experiment XII, at most sixteen units may legally be offered for sale, yet effective demand at fifty cents is twenty-two units. The *minimum* possible loss of efficiency due to the price ceiling occurs when precisely the sixteen demand units with highest redemption value are traded. The maximum attainable efficiency under the price ceiling is 95.73 percent. Whether or not this maximum is attained depends upon the rationing process. In these experiments a first come, first served method was used in which ties were broken by a random process. As can be seen on Figure 13, this rationing process induces its own inefficiencies. In every period of Experiment XII efficiency is below the 95.73 percent. Naturally other methods of solving the allocation and queueing problem resulting from the price ceiling may have different efficiency properties.

III. Summary

In summary, we found the familiar partial-equilibrium model works remarkably well to describe laboratory auction market behavior in the presence of price controls and, particularly, when the price controls are strictly binding. However, we also discovered some empirical regularities which the traditional theory cannot explain. Nonbinding price controls seem to affect the average level of prices. Furthermore, price levels and market efficiency can be influenced by removing nonbinding controls. Exactly how the standard model can be extended to explain these results is unclear. The crucial features of the institutions which induce the results have not been identified. Perhaps other institutions induce similar behavior. Perhaps many of our observations can be attributed to the single fact that institutions were *changed* and have nothing at all to do with the essential features of price controls. Nevertheless, the existence of empirical regularities seems undeniable and we offer them as a challenge to theorists who are extending the standard models to include expectations, strategic behavior, and/or the availability of market information to participants.

Subject to qualifications that must accompany any application of laboratory experimental methods, the results presented here are of potential interest to the public policy analyst. Diocletion claimed that his price ceilings would have no effect in regions where they were not binding. These results suggest that he might have been wrong. The observation that the price controls are not binding (in the sense used in partial-equilibrium analysis) is not sufficient to conclude that the controls are neutral either as to the conduct of prices or to market efficiency. Conversely, the fact that market transactions are occurring below a price ceiling or above a price floor will not be sufficient to conclude that removing controls will leave prices and quantities unchanged.

REFERENCES

Lactantius, "The Deaths of the Persecutors," in *Lactantius: The Minor Works*, translated by Sister Mary Francis McDonald, Washington 1964.

J. P. Levy, *The Economic Life of the Ancient World*, Chicago 1964.

Naphtali Lewis and Meyer Reinhold, *Roman Civilization*, Vol. II, New York 1951.

H. Michelle, *The Economics of Ancient Greece*, Cambridge 1957.

R. M. Miller, C. R. Plott, and V. L. Smith, "Intertemporal Competitive Equilibrium: An Empirical Study of Speculation," *Quart. J. Econ.*, Nov. 1977, *91*, 599–624.

C. R. Plott, "The Application of Laboratory Experimental Methods to Public Choice," in Clifford S. Russell, ed., *Collective Decision Making: Applications from Public Choice Theory*, Washington 1979.

______ **and V. L. Smith,** "An Experimental Examination of Two Exchange Institutions," *Rev. Econ. Stud.*, Feb. 1978, *45*, 133–53.

F. M. Scherer, *Industrial Pricing*, Chicago 1970.

V. L. Smith, "Experimental Economics: Induced Value Theory," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 274–79.

Leon Walras, *Elements d'Economie Politique Pure*, trans. W. Jaffe, *Elements of Pure Economics*, Homewood 1954.

A Monetary Approach to the Balance of Trade

By GARY A. CRAIG*

The predominant approach to empirical analyses of the balance of trade has been to estimate demand equations for quantities of imports and of exports, employing as explanatory variables relative prices and real incomes. This approach, referred to as the elasticities approach, is most notably exemplified by the work of Hendrick Houthakker and Stephen Magee, although other examples can be found discussed in the survey of such work by Magee (1975). In contrast, contemporary work falling under the rubric of the monetary approach to the balance of payments has emphasized that the balance of payments is determined by the net excess supply or demand for money. But because the balance of payments is identically the sum of the balance of trade, the capital account, and the service account, the monetary approach additionally implies that these subaccounts must be influenced in some way by the net excess supply or demand for money.

The empirical work of Pentti Kouri and Michael Porter and of Kouri could be interpreted as combining in one model elements of both the elasticities approach to the balance of trade and the monetary approach to the balance of payments. In these studies, the current account (the sum of the balance of trade and service account) is assumed to be exogenous to the model while the behavior of the capital account is determined by the net excess supply or demand for money. In such a framework the balance of trade could be viewed as being determined by relative prices and real incomes, and the capital account could be viewed as that component of the balance of payments which is determined by monetary factors. There is, though, a problem with such an interpretation.

In order to illustrate the difficulty, assume that beginning from an initial position of equilibrium, an excess supply of money is exogenously introduced into an economy. If that excess is assumed to be disposed of merely by the purchase of foreign assets through the capital account, the excess supply of money has been simply transformed into an excess supply of other financial assets, and a situation of disequilibrium remains. This problem can be conceptually resolved by aggregating all assets and noting that an excess supply of assets must, by Say's Law, result in a net excess demand for goods. Given exogenously the level of domestic output and prices, that net excess demand for goods must result in a trade balance deficit. It can then be concluded that, in a fixed exchange-rate regime, where prices are given exogenously, the balance of trade must be influenced by an excess supply or demand for money.

A logically consistent monetary model that properly addresses the aforementioned problem has been set forth by Jacob Frenkel and Carlos Rodríguez to describe the behavior of all of the balance of payments subaccounts. The present work empirically tests a monetary model for the balance of trade similar to theirs by developing a testable equation and then estimating it for a number of industrial countries in the postwar period. This work differs significantly not only from the empirical work on the elasticities approach to the balance of trade, but also from the empirical work on the monetary approach to the balance of payments contained in the volume edited by Frenkel and Harry Johnson. That work on the monetary approach is conducted mostly under the assumption of instantaneous equilibrium while the model to be employed herein permits adjustment over time to disequilibrium situations.

It might also be noted before proceeding that this study is conducted for a regime of fixed exchange rates while in contrast

*Assistant professor, Graduate School of Business, Columbia University. This paper is based upon a portion of my Ph.D. dissertation; for helpful comments I would like to thank the members of my thesis committee, Jacob Frenkel, Michael Mussa, and Larry Sjaastad.

the current monetary system is one of a "managed float." Despite this fact, the study is still of relevance for the current situation since, to the extent that reserve flows do occur and exchange rates are not freely flexible, elements of a fixed exchange-rate regime are introduced into the system. The current system could perhaps be best described as a weighted average of the rigidly fixed exchange-rate regime to be studied here and of a freely flexible exchange-rate regime. Also, as will be subsequently seen, the model of the trade balance for the flexible rate regime is actually a special case of that for the fixed rate regime.

I. Development of the Model

The equations that comprise the model, although related to the discussion of Frenkel and Rodríguez, are actually most directly derived from the work of Arnold Zellner et al. and Rodríguez. Zellner et al. postulate and estimate for the United States the equation

$$(1)\quad C_t = Y_t^p + \gamma_1\left(M_t^s/P_t - M_t^d/P_t\right),\ \gamma_1 > 0$$

where C_t is real consumption, Y_t^p is real income, M_t^s is the supply of money, M_t^d the demand, and P_t is the price level.[1] The parameter γ_1 is analogous to an interest rate, possessing the dimension time^{-1} and converting a stock variable to a flow. Rodríguez, working with a small country one-good model that abstracts from investment, writes the trade balance as

$$(2)\qquad B_t = Y_t - C_t$$

where B_t is the trade balance in real terms and Y_t is real current income, and then combines (1) and (2) to give[2]

$$(3)\quad B_t = \left(Y_t - Y_t^p\right) - \gamma_1\left(M_t^s/P_t - M_t^d/P_t\right)$$

This equation viewed as a model of the trade balance predicts that transitory income will be positively correlated with the trade balance while an excess supply of real cash balances will be negatively correlated with it. It is assumed here that the exchange rate is fixed and that domestic prices are exogenously determined by world prices according to purchasing power parity.

Equation (3) can be used to estimate the trade balance by specifying a simple money-demand equation[3] .

$$(4)\quad M_t^d/P_t = \alpha_1 Y_t + \beta_1 i_t, \qquad \alpha_1 > 0, \beta_1 < 0$$

where i_t is the relevant rate of interest, and substituting into (3) to yield[4]

$$(5)\qquad B_t = \delta_1\left(Y_t - Y_t^P\right) - \gamma_1\left(M_t^s/P_t - \alpha_1 Y_t - \beta_1 i_t\right)$$

or, for purposes of estimation,

$$(6)\qquad B_t = \delta_1\left(Y_t - Y_t^P\right) - \gamma_1\left(M_t^s/P_t\right) + \left(\gamma_1\alpha_1\right)Y_t + \left(Y_1\beta_1\right)i_t + \varepsilon_t$$

While equation (6) is non-linear in its parameters, it is exactly identified since the number of independent variables in it is equal to the number of parameters in (5). Thus (6) can be estimated by ordinary least squares, and it constitutes the model which is to be tested. The value of the parameter γ_1 is expected to lie between zero and unity, while the implied estimates of α_1 and β_1 should reflect the income and interest rate elasticities of the demand for money and can serve as a check on the estimates of the model.

A more complete model should include the effects on consumption of an excess supply

[1] Rudiger Dornbusch and Michael Mussa have developed a model in which this version of the "real balance effect" optimally satisfies an intertemporal allocation problem.

[2] As pointed out by the referee, equations similar to (1) and (2) were also incorporated in a model studied by S. J. Prais, whose work developed from that of Jacques Polak.

[3] Since money demand may be a function of either permanent or current income, depending upon the motivation for the holding of real cash balances, both measures are given consideration in the empirical implementation. Zellner et al. use permanent income in a similar demand for money function.

[4] The coefficient δ_1 is added since the measure of transitory income to be employed is expressed as a percentage of permanent income and not as a level.

of nonmonetary assets. Although such a model will not be estimated because of a lack of readily available data, the resulting expression for the trade balance would be of the form

(7)

$$B_t = \delta_1(Y_t - Y_t^P) - \gamma_1(M_t^s/P_t - \alpha_1 Y_t - \beta_1 i_t) - \gamma_2(K_t^s/P_t - \alpha_2 Y_t - \beta_2 i_t),$$

$$\alpha_2 > 0, \beta_2 > 0$$

where K_t^s/P_t is the supply of real nonmonetary assets and $\alpha_2 Y_t + \beta_2 i_t$ is the demand.[5] In this framework, the effect of an increase in current income on the trade balance, *ceteris paribus*, is unambiguously positive because, aside from the direct effect, it increases both the demand for money and the demand for bonds, implying that absorption must increase by less in order that the additional assets demanded be accumulated. The effect of an increase in the interest rate on the trade balance is ambiguous because while it decreases the demand for money, causing the trade balance to deteriorate, it increases the demand for bonds, causing the trade balance to improve. The net effect will depend upon the relative magnitudes of β_1 and β_2.

In the case of freely flexible exchange rates, the domestic price level is no longer linked to world prices via the exchange rate as assumed in the previous discussion; instead the price level becomes freely flexible and adjusts to clear the money market. Since excess real cash balances are disposed of through a price-level increase resulting from attempted spending, they have no effect on the trade balance, and γ_1 in (7) can be set equal to zero. Thus as noted the model of the trade balance for freely flexible exchange rates is subsumed in the model for rigidly fixed exchange rates.

II. Empirical Results

The more simple model equation (6) was estimated with annual data from *International Financial Statistics* for a number of major industrial countries roughly over the period 1953 to 1970. The starting point was determined by data availability while the endpoint was dictated by the fact that 1971 marked the beginning of a new exchange-rate system. The money supply data employed were beginning of period values, ensuring that the money supply variable is exogenous to the current period trade balance, and were defined as M_1 or M_2 depending upon which produced the better fit.[6] The interest rate variable employed was the annual average U.S. Treasury bill rate since domestic treasury bill rates were not readily available for most countries.[7] The domestic consumer price levels were used to deflate nominal values on the somewhat stringent assumption that the countries tested are small enough for that variable to be exogenous and given by the world price level. The deviations of current income from trend, computed as the residuals from a regression of the logarithm of real income on time and denoted as $(Y_t - \bar{Y})$, were employed as a proxy for transitory income. The real trade balance was constructed as the difference between exports and imports (cif), deflated by the consumer price level, while real income was measured by nominal *GNP* deflated by the consumer price level. The ordinary least squares results of the estimation of equation (6) are presented in Table 1.[8]

The results are characterized for the most part by a negative coefficient between zero and unity on the real money supply, a positive coefficient on real income, and a negative coefficient on the interest rate, as expected a priori. The money coefficients for Belgium, France, Switzerland, and the U.K. are significant at the 95 percent confidence

[5] The implied form of the consumption function underlying (7) is of course

$$C_t = Y_t^P + \delta_1(M_t^s/P_t - M_t^d/P_t) + \delta_2(K_t^s/P_t - K_t^d/P_t)$$

[6] One possible justification for the procedure used for choosing the definition of money, as suggested by the referee, is that it is a means of selecting the monetary aggregate for which the demand for money function is the most stable.

[7] For the U.K. the domestic treasury bill rate was employed as an explanatory variable.

[8] Equation (6) was also estimated by replacing current income with permanent income as an argument in the demand for money function. Permanent income was computed as $Y_t^P = fY_t + (1-f)(1+g)Y_{t-1}^P$, where $f = 0.35$ and g is the estimated rate of growth of trend income. With the exception of Denmark and Sweden, the results do not differ greatly for any of the countries.

TABLE 1—$B_t = a_1 + a_2(Y_t - \overline{Y}) + a_3(M_t/P_t) + a_4 Y_t + a_5 i_t + \varepsilon_t$
ORDINARY LEAST SQUARES

Country	a_1	a_2	a_3	a_4	a_5	F	R^2/N	$S.E./g$	$D.W./\rho$
Austria	0.142	−0.282	−0.882	0.0755	0.00294	20.37	0.87	0.0198	1.98
	(4.02)	(−1.88)	(−2.10)	(0.92)	(0.34)	(4,12)	17	0.0550	0.02
Belgium	0.233	0.681	−0.286	0.0460	0.0524	6.98	0.70	0.0533	2.36
	(2.43)	(2.22)	(−3.72)	(1.67)	(2.54)	(4,12)	17	0.0672	−0.56
Denmark	0.0203	−0.101	−0.153	−0.0290	−0.00170	22.39	0.86	0.0080	2.17
	(1.73)	(−1.60)	(−0.53)	(−0.41)	(−0.52)	(4,13)	18	0.0464	—
France	−0.209	−1.36	−0.317	0.166	−0.0197	29.72	0.91	0.0160	2.01
	(−7.16)	(−6.28)	(−8.20)	(8.42)	(−3.31)	(4,12)	17	0.0569	−0.50
Germany	0.0461	−0.716	−0.559	0.0988	−0.00213	10.86	0.78	0.0310	1.67
	(0.56)	(−3.08)	(−1.26)	(1.65)	(−0.20)	(4,12)	17	0.0660	0.65
Italy	−0.0257	−0.217	−0.085	0.101	−0.000194	6.98	0.72	0.0030	2.37
(1954–70)	(−1.74)	(−5.39)	(−0.83)	(1.05)	(−0.14)	(4,11)	16	0.0574	−0.35
Japan	−0.0136	−0.101	−0.218	0.0741	0.00121	8.09	0.75	0.0023	2.16
(1954–70)	(−5.78)	(−3.73)	(−1.85)	(1.95)	(1.10)	(4,11)	16	0.0946	−0.17
Netherlands	−0.0566	−0.255	−0.436	0.230	−0.00508*	2.79	0.48	0.0107	1.96
	(−2.25)	(−2.17)	(−1.46)	(1.42)	(−1.13)	(4,12)	17	0.0555	0.49
Sweden	−0.0412	−0.0854	−0.154	0.0757	−0.00178*	3.40	0.53	0.0046	2.27
	(−3.45)	(−1.17)	(−1.61)	(2.42)	(−0.98)	(4,12)	17	0.0415	0.34
Switzerland	−0.0619	−0.298	−0.307	0.344	−0.000449	35.99	0.92	0.0042	2.12
	(−3.23)	(−5.07)	(−4.89)	(3.98)	(−0.25)	(4,12)	17	0.0566	0.71
U.K.	0.0152	−0.124	−0.213	0.0182	0.000127	2.65	0.47	0.0024	2.08
	(1.57)	(−2.38)	(−2.53)	(1.07)	(0.18)	(4,12)	17	0.0324	−0.17

Note: The variable B_t denotes the real trade balance, Y_t real income, $\overline{Y}$ real trend income, M_t the money supply, P_t the price level, and i_t the interest rate (* indicates one-period lag). F is the F-statistic of the regression, R^2 the coefficient of determination, N the number of observations, $S.E.$ the standard error, g the estimated rate of growth of trend income, $D.W.$ the Durbin-Watson statistic, and ρ the estimated autocorrelation coefficient; t-statistics of coefficients are given in parentheses. The definition of money employed for Austria, Belgium, Denmark, Germany, and Japan was M_1 while for the remainder it was M_2. The data period is 1953 to 1970 unless noted otherwise; the source of data is *International Financial Statistics.*

level while those for Austria, Japan, and Sweden are significant at the 90 percent confidence level. The inverse of the money coefficient measures the mean time lag required for transitory real cash balances to be disposed of through the trade balance (i.e., the average length of time an increment to real cash balances is held before being exchanged for goods) and in many cases it seems extraordinarily high. Peter Jonson however found a similar result in his study of the U.K. in the sense that the mean time lag for real cash balances to be disposed of through expenditure is approximately seven and one-half years, as the analogue of a_3 (i.e. $-\gamma_1$) in his model is estimated to be -0.133. He also cites other research that has found similar seemingly extraordinarily long lags of adjustment.[9]

For all countries except Belgium the transitory income variable is, in general, highly and negatively significant, contrary to prior expectations. While the one-good consumption model predicts that transitory income will be positively correlated with the trade balance, the presence in actuality of investment goods might be sufficient to cause the marginal propensity to absorb to exceed unity and to produce a negative correlation. Some difficulty is also experienced in estimating the interest rate coefficients, and this may be due to the fact that an error in the variables problem is introduced by employing the *U.S.* rate of interest for all countries. By the interest rate parity theorem, in a regime of fixed exchange rates all interest rates will move synchronously only if the forward exchange rate is always equal to the spot exchange rate. Only with this condition, which is generally not satisfied empirically, would the use of a single rate of interest for all countries not introduce such a problem.

[9] See Jonson, p. 1003. It is also possible that the estimated value of the coefficient a_3 is biased due to the omission of the variable K_t^s/P_t in the regressions.

TABLE 2—$B_t = a_1 + a_2(Y_t - \bar{Y}) + a_3(M_t/P_t) + a_4 Y_t + a_5 i_t + \varepsilon_t$
SEEMINGLY UNRELATED REGRESSIONS

Country	a_1	a_2	a_3	a_4	a_5	a'_4	a'_5
Austria	0.119	−0.150	−0.784	0.0628	0.00319	0.388	—
	(4.41)	(−1.49)	(−3.13)	(1.30)	(0.52)		
Belgium	0.449	0.586	−0.349	0.0646	0.0480	0.502	—
	(4.26)	(2.90)	(−7.65)	(3.73)	(3.42)		
Denmark	0.0164	−0.125	−0.223	0.00421	−0.00336	0.0756	−0.250
	(1.37)	(−3.44)	(−1.48)	(0.11)	(−1.27)		
France	−0.351	−1.25	−0.341	0.181	−0.0215	1.307	−0.132
	(−14.09)	(−14.78)	(−19.39)	(18.62)	(−4.77)		
Germany	0.00210	−0.564	−0.00987	0.0239	0.000688	—	—
	(0.082)	(−4.09)	(−0.038)	(0.64)	(0.092)		
Italy	−0.0410	−0.212	−0.105	0.126	−0.000293	1.760	−0.0361
	(−5.59)	(−15.19)	(−2.73)	(3.56)	(−0.30)		
Japan	−0.0146	−0.0979	−0.230	0.0725	0.00101	1.068	—
	(−9.30)	(−9.57)	(−4.83)	(4.92)	(1.34)		
Netherlands	−0.0299	−0.323	−0.510	0.259	−0.00363	1.069	−0.0680
	(−2.78)	(−4.49)	(−3.25)	(3.00)	(−1.44)		
Sweden	−0.0304	−0.128	−0.113	0.0715	−0.00331	1.831	−0.233
	(−4.99)	(−3.68)	(−3.12)	(5.14)	(−3.12)		
Switzerland	−0.0192	−0.284	−0.306	0.358	−0.00208	1.106	−0.0358
	(−4.12)	(−8.84)	(−10.27)	(7.89)	(−1.97)		
U.K.	0.0171	−0.146	−0.219	0.0271	0.0000170	0.325	—
	(1.82)	(−4.00)	(−3.72)	(2.16)	(0.035)		

Note: a'_4 denotes the implied income elasticity of the demand for money, evaluated at the sample mean, and a'_5 the implied interest rate elasticity, also evaluated at the sample mean. All other variables are as defined in Table 1. The data period is 1954 to 1970.

It is possible to reduce the standard error of the estimated coefficients if the error terms across countries are correlated by employing the seemingly unrelated regression (*SURE*) technique developed by Arnold Zellner. Since for most countries the errors were as previously seen correlated over time, the data for them were first transformed using the estimated autocorrelation coefficients from the ordinary least square regressions. This procedure, which has been shown by Jan Kmenta and Roy Gilbert to be an efficient one, was applied to all of the countries for the period 1954 to 1970, and the results are presented in Table 2. The standard errors of the money coefficients are greatly reduced by *SURE*, and only the money coefficients of Denmark and Germany remain insignificant. The results for Germany however are somewhat puzzling since they change dramatically under *SURE*, and the coefficient on money becomes (in absolute value) very small and insignificant. A second puzzling feature of the results is the preponderant number of positive interest rate coefficients, and in particular the significance of that for Belgium and the near significance of that for Japan. The negative previous period interest rate coefficients for the Netherlands and Sweden may indicate that estimation of this coefficient could be enhanced with more appropriate data.

The last two columns of Table 2 list where deemed relevant the income and interest rate elasticities of the demand for money, evaluated at the sample means, that are implied by the regression results. On the basis of previous studies of the demand for money for the United States, the income elasticities would be expected to lie roughly in the range of 0.5 to 1.5 while the interest rate elasticities would be expected to lie roughly in the range of −0.3 to 0. With the exception of the income elasticity for Denmark, the computed values are generally consistent with prior expectations. The implied positive interest rate elasticities found for some countries, possibly due to the difficulty of estimation previously discussed, are not meaningful and consequently are not displayed.

III. Implications of the Results

An important implication of the foregoing empirical approach to the trade balance is that there is a "natural" trade balance, measured by the constant terms, most of which are significantly different from zero, around which the current trade balance varies. This implication results from the fact that, in the long run, excess real cash balances are eventually exchanged abroad for goods, resulting in the equation of money supply to money demand, while current income by definition returns to trend income. When the two explanatory variables of equation (5) are at their steady-state values of zero, the trade balance is at its natural value. As demonstrated theoretically by David Gale, there is no need for the natural value to be zero, the intuitive reasoning he cites as being essentially that one country could be on net the owner of another country's consols, which would result in a perpetual flow of goods from one to the other, equal in value to the debt service.

A second important implication of the model is that to the extent that the real cash balance effect is significant, devaluation will have an effect on the balance of payments via the trade balance. This implication results from the fact that devaluation is simply another source of an exogenous increase in the domestic price level and will impose a capital loss on the holders of real cash balances, causing through the second term on the right-hand side of equation (5) an increase in the trade balance.[10] Note that, to account for devaluation in this model, no additional modeling is required, for at least to the extent that devaluation is unexpected its effects work through diminishing the supply of real cash balances while to the extent that it is expected devaluation should be reflected in an increase in the forward exchange rate. By the interest rate parity theorem, this increase, given the foreign (world) interest rate, causes an increase in the domestic interest rate, a corresponding decrease in the demand for money, and an increase in the trade balance. This point is not unimportant since there has been some controversy in the literature as to both the theoretical and empirical existence of an effect of devaluation on the trade balance through monetary channels.[11]

A final point of interest is that the results of the estimation of the traditional import demand function can be consistent with the predictions of the elasticities approach even if the monetary approach more correctly describes the actual behavior of the trade balance. The reason is that, in the presence of nontraded goods, an increase in the money supply, *ceteris paribus*, causes through the real cash balance effect an excess demand for both traded and nontraded goods. The excess demand for traded goods is eliminated as previously seen by a temporary increase in imports and decrease in exports, or an increased trade balance deficit. The excess demand for nontraded goods in contrast is eliminated by a temporary price increase of nontraded goods relative to traded goods, where the prices of the latter are assumed to be given on world markets. Thus a monetary expansion produces an inverse correlation between the quantity of imports and the relative price of traded to nontraded goods and thus between the quantity of imports and the price of imports relative to the general price level. That relation then could be estimated and mistakenly interpreted as a demand for imports function. The argument does not completely dismiss the results of the elasticities approach, however, since the success of the estimation of the export demand function cannot be similarly explained, the price variable for that function usually being the price of one country's exports relative to a weighted average of the export prices of all other countries.

[10] Connolly and Taylor (1976b) present evidence on the upward movement of the price level in developing countries following devaluation.

[11] See the literature reviewed by Magee (1976), pp. 166–67.

IV. Conclusion

The empirical results presented tend to confirm the argument that under a fixed exchange-rate regime monetary factors have a direct influence on the balance of trade. Holding constant the effects of deviations of current from trend income, an excess supply of real cash balances has as expected a negative effect on the balance of trade. However the prediction of the model that an increase in transitory income will result in an increase in the balance of trade is contradicted by the evidence. Holding constant the effects of money supply and demand, an increase of current over trend income has contrary to expectations a negative effect on the balance of trade. The fact that this aspect of the monetary model is in conflict with the evidence might be due either to the omission in the analysis of an explicit distinction between investment and consumption goods or of one between traded and nontraded goods. Further consideration of these elements in a more complete model may resolve this inconsistency of the theory with the evidence.

REFERENCES

M. Connolly and D. Taylor, (1976a) "Devaluation, Traded Goods, and International Assets," in E. Claassen and P. Salin, eds., *Recent Developments in International Monetary Economics*, Amsterdam 1976.

______ **and** ______ (1976b) "Testing the Monetary Approach to Devaluation in Developing Countries," *J. Polit. Econ.*, Aug. 1976, *84*, 849–59.

R. Dornbusch and M. Mussa, "Consumption, Real Balances and the Hoarding Function," *Int. Econ. Rev.*, June 1975, *16*, 415–21.

Jacob A. Frenkel and Harry G. Johnson, *The Monetary Approach to the Balance of Payments*, Toronto 1976.

______ **and Carlos A. Rodríguez,** "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

D. Gale, "The Trade Imbalance Story," *J. Int. Econ.*, May 1978, *4*, 119–37.

H. S. Houthakker and S. P. Magee, "Income and Price Elasticities in World Trade," *Rev. Econ. and Statist.*, May 1969, *51*, 111–25.

P. D. Jonson, "Money and Economic Activity in the Open Economy: The United Kingdom, 1880–1970," *J. Polit. Econ.*, Oct. 1976, *84*, 979–1012.

J. Kmenta and R. F. Gilbert, "Estimation of Seemingly Unrelated Regressions with Autoregressive Disturbances," *J. Amer. Statist. Assoc.*, Mar. 1970, *65*, 186–97.

P. J. K. Kouri, "The Hypothesis of Offsetting Capital Flows," *J. Monet. Econ.*, Jan. 1975, *1*, 21–40.

______ **and Michael G. Porter,** "International Capital Flows and Portfolio Equilibrium," *J. Pol. Econ.*, May/June 1974, *82*, 443–68.

S. P. Magee, "The Empirical Evidence on the Monetary Approach to the Balance of Payments and Exchange Rates," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 163–70.

______, "Prices, Incomes, and Foreign Trade," in Peter Kenen ed., *International Trade and Finance: Frontiers for Research*, Cambridge 1975.

J. J. Polak, "Monetary Analysis of Income Formation and Payments Problems," *Int. Mon. Fund Staff Papers*, Nov. 1957, *6*, 1–50.

S. J. Prais, "Some Mathematical Notes on the Quantity Theory of Money in an Open Economy," *Int. Mon. Fund Staff Papers*, May 1961, *8*, 212–26.

C. A. Rodríguez, "The Terms of Trade and the Balance of Payments in the Short Run," *Amer. Econ. Rev.*, Sept. 1976, *66*, 500–23.

A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *J. Amer. Statist. Assoc.*, June 1962, *57*, 348–68.

A. Zellner, D. S. Huang, and L. C. Chau, "Further Analysis of the Short-Run Consumption Function with Emphasis on the Role of Liquid Assets," *Econometrica*, July 1965, *33*, 571–81.

International Monetary Fund, *International Financial Statistics*, Washington, various issues.

On Nonbinding Price Controls in a Competitive Market

By VERNON L. SMITH AND ARLINGTON W. WILLIAMS*

Interest in the effect of nonbinding price controls on double auction markets stems from two primary considerations. The double auction institution converges to a competitive allocation more rapidly, and with fewer participating agents than any other institution with which it has been compared (see Smith et al.). One way to improve our understanding of this important property is to determine what conditions, if any, can interfere with or retard this convergence process. Nonbinding price controls represent a condition that may affect this convergence process. Hence, if such effects can be documented, they will provide a body of data that any future proposed model of the double auction process should be able to explain. A second reason for studying the effect of nonbinding controls on the double auction is practical: The organized commodity exchanges "...often set limits on price fluctuations during any single day. When prices at any point during a day rise above ·ǫr fall below the closing prices of the preceding day by more than the amount of the limit, no further trading for that day is permitted" (Walter Labys, p. 162). Consequently, commodity trading frequently occurs at prices near the level of nonbinding price floors or ceilings.

Mark Issac and Charles Plott report the results of twelve exploratory experiments in which various price control constraints are imposed on double auction markets. Their two principal conclusions can be summarized as follows:

1) The hypothesis is rejected that nonbinding price controls, that is, price ceilings above or price floors below the competitive equilibrium (*CE*), will serve as a focal point or signalling price on which sellers and buyers will key their contracts.

2) Inconclusive evidence is presented in support of the hypothesis that nonbinding controls near the *CE* will bias prices *below CE* when there is a price ceiling and *above CE* when there is a price floor.

Support for this second hypothesis is not conclusive because some experimental markets show a tendency to converge from below and others from above depending upon the relative bargaining strength of buyers as against sellers. Thus sampling variation among subjects can yield a group in which buyers (sellers) are able to make contracts at an average price below (above) *CE* for several periods of trading. Consequently, in an experiment in which there is a price ceiling (floor) five cents above (below) *CE* and in which contract prices are observed to occur below (above) *CE*, one cannot determine conclusively whether the observed effect was due to the nonbinding price control or to the bargaining characteristics of the market participants.

We report below an experimental design developed for the purpose of separating these confounding factors and allowing the effect of nonbinding controls to be isolated. The results of sixteen experiments strongly support the hypothesis that markets with a nonbinding price ceiling (floor) near *CE* will converge from below (above) relative to any otherwise identifiable tendency to converge from below (above). An analysis of the effect of a nonbinding price ceiling (floor) on the distributions of bids and offers reveals the cause of this bias: ceilings limit the bargaining strategies of sellers especially, but also that of buyers, while floors have the opposite effect. Thus, in the absence of price controls, double auction trading is characterized by a process in which sellers typically make concessions from offer prices well above *CE* while buyers most often concede from bid prices well below *CE*. A price ceiling truncates seller offer prices at the ceiling, requir-

*University of Arizona and Indiana University, respectively. Research support by the National Science Foundation is gratefully acknowledged.

ing them to begin their bargaining from a less advantageous position at the ceiling or below. Buyer bids are also effected, but less dramatically, in that the occassional bid above *CE* that might occur in a free market, will be blocked by the ceiling. Also the ceiling, and/or the consequent lower offers by sellers, induces somewhat lower-bidding behavior by buyers.

I. Experimental Design

Since the research task is to isolate the treatment effect of price controls on competitive market dynamics, and since this effect can be obscured by noise and confounding factors, we have devoted considerable care to the development of an appropriate experimental design. The resulting design has the following principal characteristics:

1) All experiments used the PLATO computer version of the double auction exchange mechanism developed by Williams (1980). The computer permits better control over "experimenter" effects by assuring uniform procedures across all experiments, with accurate computerized recording of all bids, offers, contracts, and their time of occurrence. The particular form of the double auction used in the price control experiments employs both the New York Stock Exchange "improvement rule", and a computerized version of the "specialist book." During a particular auction sequence, the improvement rule (rules 71 and 72 on the Exchange; see George Leffler and Loring Farwell, pp. 187–88) requires a bid (offer) to be higher (lower) than an outstanding bid (offer) before it can be announced. When a trade occurs, the "auction" of the unit ends and the market (in our case PLATO) awaits new bids and offers that must provide sequentially improving terms. A bid (offer) that is lower (higher) than the outstanding bid (offer) is entered into a PLATO queue, lexicographically ordered with higher bids (lower offers) having priority, and tied bids (offers) ordered chronologically—the first in having priority over the second, etc.

2) Only subjects who had participated in at least one previous PLATO double auction market, without price controls and defined by different supply and demand parameters, were used in the experiments. (See our earlier paper for a discussion of alternative bidding rules and trading experience as experimental treatment variables in PLATO double auctions.)

3) The induced values and costs or limit prices (see Smith) for a typical experiment are shown in Table 1 for each of four buyers and four sellers. (The theoretical total surplus is $10.20 per trading period, with commissions of $3.00 per period for fifteen traded units, giving a total payout of $198 per experiment.) The corresponding market supply and demand with symmetrical buyer and seller surpluses are shown on the lower left of Figure 1. Each buyer (seller) receives a cash payment equal to the difference between his/her value (selling price) for a unit and his/her purchase price (cost) plus a ten cent commission for each unit traded. Except for the commissions, buyers (sellers) earn the realized consumer's (producer's) surplus. Notice that in this design (Figure 1) there are several intramarginal and submarginal units in the range from five cents above to five cents below the *CE* price. Hence, inefficient submarginal trades can easily occur if there are many contracts away from the *CE* price. Similarly, intramarginal units near the *CE* price are less well motivated to trade than other more profitable units, leading to an increased chance of inefficiency. These features specify a supply and demand design in which efficiency, defined as the ratio of realized to theoretical buyers' plus sellers' surplus, is likely to be sensitive to factors, such as nonbinding price controls, hypothesized to interfere with the trading process. These features were not present in the Issac and Plott (Figure 2) design in which intramarginal and submarginal units were ten cents or more above or below the *CE* price.

4) All experiments consist of three "weeks" of trading, each week consisting of five (or four in some experiments) trading periods. Week 1 provides the baseline set of observations with *no* price control. If a particular group is characterized by relatively strong bargaining buyers, this is measured by the difference between the total surplus realized by buyers and the total surplus obtained

Table 1—Unit Values and Costs in Dollars

Subject	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Individual Competitive Equilibrium Surplus
Buyer 1	5.35	5.10	4.70	4.60	4.50	–	1.20
Buyer 2	5.60	4.90	4.80	4.65	4.55	–	1.35
Buyer 3	5.60	4.90	4.80	4.65	4.60	4.50	1.35
Buyer 4	5.35	5.10	4.70	4.65	4.55	–	1.20
Seller 1	3.95	4.20	4.60	4.65	4.75	–	1.20
Seller 2	3.70	4.40	4.50	4.65	4.75	–	1.35
Seller 3	3.95	4.20	4.60	4.70	4.80	–	1.20
Seller 4	3.70	4.40	4.50	4.65	4.70	4.80	1.35
Total Market Surplus							10.20

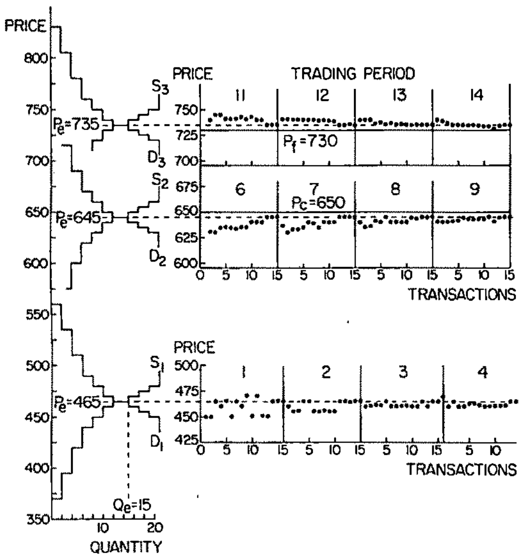


Figure 1. Experiment 2:26

by sellers. Since in our design, differential surplus is zero at the *CE*, this measure should reflect only the relative bargaining strength of buyers in any particular experiment.

5) In each experiment, following the completion of the first week of trading, a prespecified constant is added to each limit price value and cost unit, thereby uniformly shifting the supply and demand schedules up or down relative to Week 1. Also the assignment of unit values (costs) to buyers (sellers), as illustrated in Table 1, is rerandomized by reassigning the shifted buyer (seller) limit price valuations (costs) among the buyers (sellers). Trading is then resumed in Week 2, periods 6 through 9 or 10, under these new conditions. In eight experiments, a price ceiling is also imposed. In four other experiments a price floor is imposed, and in four others no price control is imposed. This procedure is designed to allow any effect of the shift in supply and demand to be separated from the effect of the price controls. Subjects are informed that trading will proceed during Week 2 under a price ceiling (at, for example $6.50 as in Figure 1) by the appearance of the following message on their display screens at the end of Week 1:

SPECIAL ANNOUNCEMENT

> Price controls will be in effect during market Week 2. The MAXIMUM allowed bid or offer price will be $6.50. Any entry which violates the above will be automatically rejected and will generate a descriptive error message.

If a subject attempts to violate the $6.50 price ceiling the following message is generated: "Your entry exceeds the maximum allowed price of $6.50." The announcement of a price floor is made in exactly the same format as given above with "MINIMUM" replacing "MAXIMUM" in the announcement's text.

6) Following the completion of the second week of trading, the valuations (costs) are again shifted and rerandomized. Trading is then resumed in Week 3, period 11 through

TABLE 2—NUMBER OF EXPERIMENTS PERFORMED UNDER EACH TREATMENT CONDITION

	Price Control Variable		
Supply and Demand Shift in Week 2 (3)	No Price Control in Week 2 or Week 3	Week 2 Price Ceiling 5 Cents Above *CE* Week 3 Price Floor 5 Cents Below *CE*	Week 2 Price Floor 5 Cents Below *CE* Week 3 Price Ceiling 5 Cents Above *CE*
Up (Up)	1	2	1
Up (Down)	1	2	1
Down (Up)	1	2	1
Down (Down)	1	2	1

14 or 15, with a price floor in eight experiments, a price ceiling in four experiments, and with no price control in four. The number of experiments conducted under each Week 2–Week 3 shift condition, with or without a price ceiling (floor), is shown in Table 2. It should be noted, however, that the design is still not completely "balanced" in that only four were conducted using a price floor in Week 2 followed by a price ceiling in Week 3. Although it would be scientifically appealing, we thought that it was perhaps not worth the cost (subject earnings are about $200 per experiment) to fill in these additional cells with experimental observations.

Based on this experimental design, we propose two linear models for separating the effect of 1) the differential bargaining strength of buyers relative to sellers, 2) a uniform shift in supply and demand with individual random reassignment of units, 3) a price ceiling just above *CE*, and 4) a price floor just below *CE*. Define:

$B(t)$: Buyer realized surplus (earnings net of commissions) in period t.

$S(t)$: Seller realized surplus in period t.

$D(t) \equiv B(t) - S(t)$: Differential bargaining strength of buyers over sellers.

$$X_i^c = \begin{cases} 1, \text{ if ceiling price is imposed} \\ \quad \text{in Week } i. \\ 0, \text{ if no price control is imposed} \\ \quad \text{in Week } i. \end{cases}$$

$$X_i^f = \begin{cases} 1, \text{ if floor price is imposed} \\ \text{in Week } i. \\ 0, \text{ if no price control is imposed} \\ \text{in Week } i. \end{cases}$$

$$Y_i = \begin{cases} 1, \text{ if supply and demand shift down} \\ \quad \text{in Week } i. \\ 0, \text{ if supply and demand shift up} \\ \quad \text{in Week } i. \end{cases}$$

The proposed linear models are stated:

(1)
$$D(t) = \begin{cases} \alpha_2 D(t-5) + \beta_2 X_2^c + \gamma_2 X_2^f + \delta_2 Y_2 \\ \alpha_3 D(t-10) + \beta_3 X_3^c + \gamma_3 X_3^f + \delta_3 Y_3 \end{cases}$$

If $\alpha_i > 0$ it means that whether buyers are stronger ($D > 0$) or weaker ($D < 0$) in a particular experimental group, this characteristic tends to persist across comparable trading periods (for example, 1, 6, and 11; 2, 7, and 12; etc.) in successive weeks. If $\alpha_3 < \alpha_2 < 1$, this suggests week-to-week learning (in the sense of continued convergence to *CE*) after correcting for the effect of shifts in supply and demand. The principle research hypothesis of this paper is that $\beta_i > 0$, and $\gamma_i < 0$, that is, a nonbinding price ceiling favors buyers by lowering contract prices relative to *CE*, while a nonbinding price floor favors sellers by raising contract prices relative to *CE*. Finally, the effect of a uniform shift in supply and demand could cause prices to overshoot the new equilibrium ($\delta_i > 0$), benefiting buyers (sellers) when there is a downward (upward) shift. Alternatively, following a supply and demand shift, prices might undershoot the new equilibrium ($\delta_i < 0$). If successive shifts in supply and demand have a diminished disequilibrating effect, then we would expect $|\delta_2| > |\delta_3|$.

II. Experimental Results

Table 3 lists the mean deviation of contract prices from the *CE* price, and the ef-

TABLE 3—MEAN DEVIATION OF CONTRACT PRICES FROM *CE* (AND EFFICIENCY) BY WEEKS

Experiment Number	Price Controls	D and S Shift	Week 1	Week 2	Week 3
2:18	Ceiling, Week 2	Up, Week 2	.058	.024	.066
	Floor, Week 3	Up, Week 3	(99.63)	(100.)	(99.51)
2:26	Ceiling, Week 2	Up, Week 2	−.036	−.056	.029
	Floor, Week 3	Up, Week 3	(99.39)	(99.51)	(99.88)
2:27	Ceiling, Week 2	Down, Week 2	−.095	−.113	−.010
	Floor, Week 3	Up, Week 3	(98.90)	(99.26)	(99.02)
2:30	Ceiling, Week 2	Down, Week 2	−.041	−.043	.061
	Floor, Week 3	Up, Week 3	(95.46)	(98.53)	(97.67)
2:35	Ceiling, Week 2	Down, Week 2	−.034	−.112	.020
	Floor, Week 3	Down, Week 3	(99.51)	(99.39)	(99.39)
2:36	Ceiling, Week 2	Up, Week 2	−.021	−.010	.030
	Floor, Week 3	Down, Week 3	(99.75)	(100)	(100)
2:40	Ceiling, Week 2	Up, Week 2	.017	−.003	.069
	Floor, Week 3	Down, Week 3	(99.26)	(99.26)	(99.63)
2:41	Ceiling, Week 2	Down, Week 2	−.031	−.026	.005
	Floor, Week 3	Down, Week 3	(99.63)	(98.65)	(99.88)
2:49	None	Down, Week 2	−.030	−.036	.004
		Up, Week 3	(100)	(99.39)	(100)
2:54	None	Up, Week 2	.042	.040	−.004
		Down, Week 3	(99.14)	(99.39)	(99.51)
2:56	None	Up, Week 2	.113	−.005	−.023
		Up, Week 3	(97.67)	(99.63)	(100)
2:57	None	Down, Week 2	.046	−.008	−.023
		Down, Week 3	(98.16)	(99.51)	(99.51)
3:8	Floor, Week 2	Up, Week 2	−.063	.063	−.069
	Ceiling, Week 3	Up, Week 3	(99.14)	(99.39)	(98.53)
3:10	Floor, Week 2	Down, Week 2	.032	.030	−.093
	Ceiling, Week 3	Up, Week 3	(99.75)	(99.88)	(99.02)
3:12	Floor, Week 2	Down, Week 2	−.142	−.001	−.093
	Ceiling, Week 3	Down, Week 3	(97.67)	(96.32)	(99.35)
3:21	Floor, Week 2	Up, Week 2	−.227	.039	−.134
	Ceiling, Week 3	Down, Week 3	(93.75)	(99.88)	(97.67)

ficiency by weeks for all sixteen experiments. These experiments, which are listed in chronological order, were conducted over a period of nearly two years and were interspersed with a large number of other double auction experiments with quite different research objectives. From the mean price deviations in Table 3 for the sixteen price control experiments, one can discern a strong tendency for prices to be lower (higher) in Week 2 when the nonbinding ceiling (floor) is in effect relative to the baseline Week 1, and higher (lower) in Week 3 when the nonbinding floor (ceiling) is in effect relative to Week 1. Mean price deviations in Week 2 are consistent with this observation for every price control experiment except 2:36, 2:41, and 3:10, while in Week 3 mean price deviations violate this observation only in experiments 3:12 and 3:21.

The effect of the uniform shift in supply and demand on price deviations relative to the baseline week is not obvious by inspection, although there appears to be a tendency for prices to overshoot the new equilibrium. Efficiency, measured by realized buyers' plus sellers' surplus as a percentage of theoretical total surplus, is close to 100 percent under all treatments, which is highly characteristic of PLATO double auction experiments using experienced subjects (see Williams; Smith and Williams). Table 3 shows some tendency for efficiency to improve across the three weeks of trading, regardless of whether there is or is not a price control, and independently of the shift condition. This suggests that week-to-week learning dominates price controls, as well as supply and demand shifts, in affecting market efficiency. A Wilcoxon test of the hypothesis that price controls lead

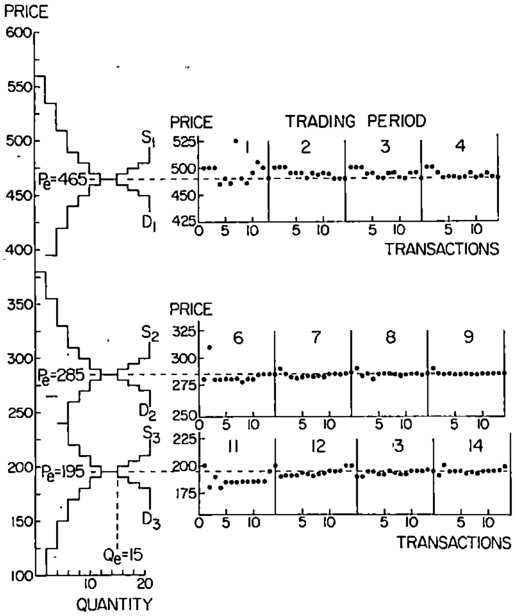


FIGURE 2. EXPERIMENT 2:57

to a relative decrease in efficiency is easily rejected.

Figures 1 and 2 display the contract price sequences by trading period for experiments 2:26 (with price controls), and 2:57 (without price controls).[1] These two experiments illustrate the tendency (measured in the regression estimates reported below) for 1) the price ceiling (floor) to cause contract prices to occur below (above) the *CE* relative to the baseline week; and 2) the shift in supply and demand to cause contract prices to overshoot the new equilibrium relative to baseline. The charts also illustrate the pronounced tendencies to converge to the *CE* under *all treatment conditions.* The effect of price controls is merely to retard this convergence, and to cause convergence to be from below (above) when there is a price ceiling (floor). Nonbinding price controls affect market dynamics, but not static equilibria.

The results of estimating the coefficients in the regression model (1) yield,

$$(2)\quad D(t)=$$

$$\begin{cases} \underset{(3.41)}{0.236}\,D(t-5)+\underset{(2.08)}{0.464}\,X_2^c-\underset{(-6.03)}{2.02}\,X_2^f \\ \qquad +\underset{(4.81)}{1.089}\,Y_2,\ R^2=.55,\ N=70 \\ \underset{(1.94)}{0.111}\,D(t-10)+\underset{(8.47)}{2.315}\,X_3^c\,-\underset{(-6.54)}{1.190}\,X_3^f \\ \qquad +\underset{(1.38)}{.260}\,Y_3,\ R^2=.74,\ N=67 \end{cases}$$

which supports the following conclusions:

1) The price ceiling transfers an average of about 46 cents per period in surplus from sellers to buyers during Week 2 and \$2.32 during Week 3. The floor transfers \$2.02 per period from buyers to sellers during Week 2 and \$1.19 during Week 3. The *t*-values shown in parenthesis indicate that the ceiling and floor regression coefficients are highly significant ($P<.025$, one-tailed test). We reject the hypothesis that nonbinding price controls near the *CE* have no effect on the dynamics of the equilibrating process in favor of the hypothesis that they are effective in the a priori predicted direction.

2) Differential bargaining strength by either buyers or sellers, in particular double auction markets, tends to persist in successive trading periods, and weeks, but with decreasing effect. The differential buyer-seller surplus in Week 2 averages 24 percent of its Week 1 level, and by Week 3 it is only 11 percent of its Week 1 level.

3) A uniform shift, up or down, in demand and supply, and a rerandomization of the assignments of individual supply and demand, causes some overshooting of the new equilibrium. A downshift in Week 2 redistributed \$1.09 of surplus from sellers to buyers relative to an upshift. In Week 3, this redistribution is reduced to about twenty-six cents. This implies that double auction markets show an improved ability to track changes in *CE* with successive changes in supply and demand. Experienced market

[1] Most of the experiments consisted of five periods of trading in any given week. Figures 1 and 2 display only the contract prices for the first four trading periods of each week. The fifth period results in most all of the experiments produced trades very near the *CE*.

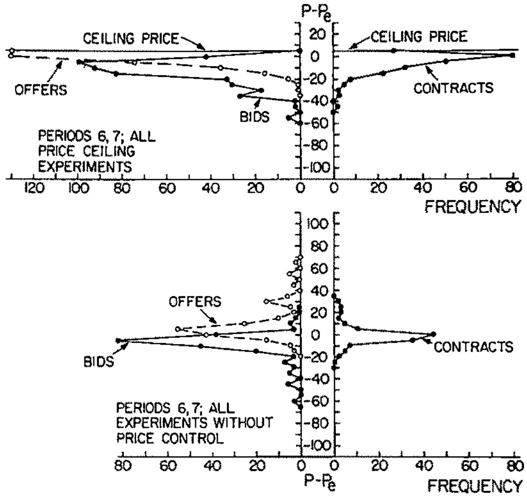


FIGURE 3. PRICE CEILING FREQUENCY DISTRIBUTIONS

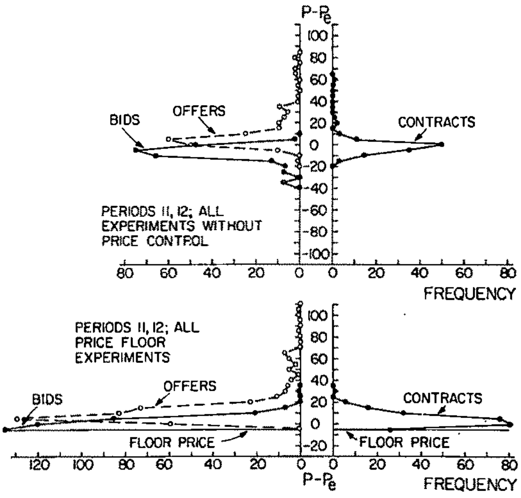


FIGURE 4. PRICE FLOOR FREQUENCY DISTRIBUTIONS

participants learn to adapt more quickly, in terms of convergence speed, to shifts in supply and demand so that this further "experience" has an identifiable treatment effect.

III. How Nonbinding Price Controls Interfere with the Double Auction Bargaining Process

Issac and Plott speculate that their conjectured effect of nonbinding price controls in biasing prices away from the equilibrium "may have something to do with information and 'search'." Our PLATO computerization of the double auction institution makes it feasible to examine the effect of price controls on the distributions of bids and offers.

The upper half of Figure 3 plots the distributions of bids and offers on the left, and the distribution of contracts on the right, pooled across all Week 2 price ceiling experiments for trading periods 6 and 7. The lower half of Figure 3 plots the corresponding distributions for periods 6 and 7 across all experiments in which no price control is in effect during Week 2. The most obvious effect of the price ceiling is to truncate the bid and offer distributions above the ceiling. But this truncation is most pronounced with the offer distribution as sellers are required to begin their bargaining with offers at or below the ceiling. One would expect experienced buyers rarely to enter bids as high as the ceiling, even in the absence of the ceiling.

Similarly, a price floor five cents below the *CE* price truncates the bid and offer distributions from below, but it is the bid distribution that is most strongly affected. This is seen in Figure 4 comparing the bid, offer and contract distributions without a floor with the corresponding distributions when a floor is in effect.

Table 4A displays a comparison of the mean price quotations generated under the ceiling (Week 2)–floor (Week 3) treatment sequence with those generated in the experiments with no price controls. The mean bid and the mean offer are lowered by a price ceiling, but the decrease in the mean offer is much larger than for the decrease in mean bid. A price floor reverses this effect, with the mean bid increasing more than the mean offer increases. The Mann-Whitney unit normal deviate (Z_u) indicates rejection of the hypothesis that the period 6–7 offer distributions are identical with and without a price ceiling. This hypothesis is also rejected for the period 6–7 bids, but at a lower significance level. The same qualitative conclusions hold when comparing period 11–12 bids (offers) with and without a price floor.

Table 4B compares mean price quotations generated under the floor (Week 2)–ceiling (Week 3) treatment sequence with those gen-

TABLE 4A—MEAN DEVIATIONS OF QUOTATIONS FROM *CE* AND MANN-WHITNEY TESTS OF SIGNIFICANCE

Price Control Condition	Quotation	
	Bids	Offers
Trading Periods 6–7		
Price Ceiling		
Experiments	−0.170	−0.015
No Price Control		
Experiments	−0.110	0.176
Mann-Whitney, Z_u	6.63	11.0
Trading Periods 11–12		
No Price Control		
Experiments	−0.099	0.113
Price Floor		
Experiments	0.003	0.137
Mann-Whitney, Z_u	12.2	4.25

TABLE 4B—MEAN DEVIATIONS OF QUOTATIONS FROM *CE* AND MANN-WHITNEY TESTS OF SIGNIFICANCE

Price Control Condition	Quotations	
	Bids	Offers
Trading Periods 6–7		
Price Floor		
Experiments	.001	.238
No Price Control		
Experiments	−.110	.176
Mann-Whitney, Z_u	10.43	5.80
Trading Periods 11–12		
No Price Control		
Experiments	−.099	.113
Price Ceiling		
Experiments	−.464	−.059
Mann-Whitney, Z_u	13.99	12.22

erated without price controls. As in Table 4A, the values of Z_u indicate rejection of the hypothesis of identical distributions in all cases. However, the Week 3 price ceiling has a much stronger effect on the bid distribution than was the case with the Week 2 ceiling (Table 4A). This can be explained, at least partially, by noting that three of the four experiments run under the floor-ceiling sequence were characterized during Week 1 trading as having relatively "strong" buyers (as indicated by the negative mean price deviations from *CE* given in Table 3). In contrast, three of the four experiments run without price controls display positive mean price deviations during Week 1, indicating that sellers were somewhat stronger than buyers. The comparisons of price quote distributions across only the price control treatment conditions do not control for the relative bargaining strengths displayed by each subject group.

These data show that price ceilings or floors tend to interfere asymmetrically with the double auction bargaining process. Price ceilings limit the bargaining strategies of sellers. With a ceiling price, sellers must learn to refrain from making competitive offer concessions much below the ceiling, or avoid accepting bids until buyers are bidding near the ceiling. But a price ceiling also lowers the bids of buyers, partly because the relatively rare high bids are truncated by the price ceiling, and partly perhaps because the buyers learn that, with the ceiling, they can induce the sellers to accept somewhat lower bids. These considerations also apply when there is a price floor, except that the position of buyers and sellers, and the directional effects are reversed.

REFERENCES

R. M. Issac and C. R. Plott, "Price Controls and the Behavior of Auction Markets: An Experimental Examination," *Amer. Econ. Rev.*, June 1981, *71*, 448–59.

W. C. Labys, "Bidding and Auctioning on International Commodity Markets." in *Bidding and Auctioning for Procurement and Allocation*, New York 1976.

George L. Leffler and Loring C. Farwell, *The Stock Market*, New York 1963.

V. L. Smith, "Experimental Economics: Induced Value Theory," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 274–79.

______ **and Arlington W. Williams,** "An Experimental Comparison of Alternative Rules for Competitive Market Exchange," Yale Univ. Conference on Auctions and Bidding (Dec. 1979), rev. Mar. 1980.

______ **et al.,** "Computerized Competitive Market Institutions: Double Auctions versus Sealed-Bid Auctions," Univ. Arizona, rev. Apr. 1981.

A. W. Williams, "Computerized Double Auction Markets: Some Initial Experimental Results," *J. Bus. Univ. Chicago*, July 1980, *53*, 235–58.

Firm-Specific Human Capital as a Shared Investment

By MASANORI HASHIMOTO*

The standard analysis of firm-specific human capital argues that the cost of and the return to the investment will be shared by the worker and the employer. By sharing the investment, the parties reduce the likelihood of either party unilaterally terminating the employment relationship and imposing on the other party a loss in his return. This argument, originally advanced by Gary Becker (pp. 10–15), has become accepted almost as a theorem.[1] The sharing decision is particularly important in determining the shape of the wage profile and the behavior of labor turnover in the labor market. The exact decision process involved in determining the sharing arrangement, however, appears to have received little attention in the literature.[2]

In this paper, a formal statement of the sharing model is presented. The model allows a systematic analysis of the incentive to share the investment in firm-specific human capital. My analysis reveals that whether or not the investment is shared depends on the existence in the post-investment years of costs of evaluating and agreeing on the worker's productivities in the firm and elsewhere.

This paper and two others (my 1979 article and my article with Ben Yu) demonstrate the usefulness of the sharing model. My 1979 article develops and tests the hypothesis that the ubiquitous bonus payments in Japan can be understood as payments for the returns to firm-specific human capital. The paper with Yu extends the analysis of firm-specific human capital by considering the incentives for introducing wage flexibility in employment contracts. Various dismissal and quitting rules are also compared in that paper.

In this paper, I use the model in its simplest form to offer a formalization of Becker's hypothesis concerning the sharing of the gains and costs of specific training. In so doing, I demonstrate that the Becker hypothesis can be viewed as a direct application of the Coase Theorem. Implications of the model for the experience-earnings profile are also discussed. This paper also clears up some confusion in the literature about the validity of the sharing hypothesis (see fn. 1).

I. A Model of Specific Investment

According to the standard analysis, the worker invests in specific human capital by accepting a wage lower than his alternative wage, and receives a return on his investment during the post-investment periods in the form of a wage higher than his alternative wage. The employer invests in specific human capital by paying the worker a wage larger than the value of his marginal product, and receives a return on the investment in subsequent periods by paying a wage smaller than the value of his marginal product.

In the standard analysis, the motivation for sharing the investment is said to be the uncertainty about the parties' post-investment behavior which affects the

*University of Washington. I wish to acknowledge my debt to Yoram Barzel, George Borts, Carl Cutter, Ben Yu, and an anonymous referee for valuable comments on earlier versions of this paper. Any remaining shortcomings are my responsibility.

[1] Recently, David Donaldson and B. Curtis Eaton attempted to dispute Becker's sharing hypothesis. As the subsequent comment by Sheila Eastman pointed out, however, their analysis is based on a definition of investment different from the conventional definition, and thus fails to properly dispute the sharing argument.

[2] For example, interesting studies by Donald Parsons (1972) and John Pencavel were concerned more with the implications of investment in specific human capital on labor turnover than with the determination of the sharing arrangement. Dale Mortensen's interesting paper assumes zero transaction costs and effectively rules out the incentive to share the investment. See the related discussion in fn. 7. See, also my 1979 article, which uses a related model to the one presented here to analyze bonus payments in Japan. Finally, Parsons (1977) provides an informative review of the recent literature on related research.

capturability of the returns to the investment. Here it is argued that the motivation is not uncertainty per se but the existence of transaction costs. In the post-investment period, the parties face positive transaction costs with the result that the employer may dismiss the worker or the worker may quit, even though there may be a net loss from a separation. To minimize the loss from "non-optimal" separations, the parties determine the optimum sharing prior to undertaking the investment. The following model formalizes this decision process by applying the proposition made by Ronald Coase to the context of the income streams generated by specific training. The analysis below explicitly incorporates the sharing decision in the investment model. In this way, the model will help determine the extent of sharing under different situations, and thus potentially permits a test of the hypothesis.

Let us consider a two-period model, and assume that both the worker and the employer are risk neutral, and that both the capital and labor markets are perfect. Suppose that employment in a given firm entails investment in firm-specific human capital in the first period. The employer and the worker first decide on the amounts of investment and how to share the investment. At the beginning of the second period, all relevant facts become known, and the parties decide whether to stay together or to separate. The investment entails costs in terms of foregone output caused by the worker having to spend some working hours on producing human capital and other expenses. Assume that the worker possesses H units of completely general human capital to start with. Given a production function for human capital and the prices of inputs, a cost function for human capital can be derived as

$$C=C(h) \quad C'>0, C''>0 \tag{1}$$

where C is cost and h is the amount of human capital to be produced. Although the initial human capital may affect the cost function, we ignore this effect for simplicity's sake.

The value of product per unit of h is m in this firm, and zero in the alternative employment. In other words, h is completely specific to this firm. Assume for the moment that the parties agree costlessly on the value of m. This assumption will be relaxed subsequently. In the second period, the value of the marginal product of the worker in this firm is given by

$$v=H+mh \tag{2}$$

where the value of initial general human capital is assumed to be unity in this and in the alternative employment. Thus, the alternative value of the worker is given by

$$y=H \tag{3}$$

in both periods. Then, the return to investment, R, will be

$$R=v-y=mh \tag{4}$$

The worker's second-period wage, w, may be viewed as consisting of alternative wage and a wage premium representing his share of the return to the investment.[3] That is,

$$w=y+\alpha R \quad 0\leqslant\alpha\leqslant 1 \tag{5}$$

where α is the worker's sharing ratio in the returns to the specific human capital. The employer's return from the investment will be

$$r=v-w=(1-\alpha)R \tag{6}$$

Suppose that the investment turns out to be an error *ex post*. Given this contract, any separation decision is optimal regardless of the value of α. this point is easily seen by recognizing that the criteria for quit, $w-y\leqslant 0$, and for dismissal, $r\leqslant 0$, imply $v\leqslant y$, or that the alternative value is at least as large as the value in this firm. If $\alpha=0$, then $w=y$, and the worker is indifferent given the choice of separation, but the employer is not. If $\alpha=1$, then $r=0$, and the employer is indifferent given the choice of separation, but the worker is not. In either case, one of the

[3] The worker's first-period wage will be determined later from the long-run equilibrium conditions.

parties makes the correct decision for both parties. If α is between zero and unity, both parties have the incentive to separate only when $v<y$. The value of α is irrelevant to the separation decisions because the parties are assumed to transact costlessly on the values of v and y. Put another way, although α assigns the property rights to the investment returns, with zero costs of transacting the assignment of such rights to either party leaves the efficiency of the ultimate use of resource unaffected. This implication is nothing more than a statement of the Coase Theorem.

Clearly then, the investment is shared because of the existence in the post-investment period of costs of evaluating and agreeing on the values of v and y. These transaction costs introduce important sources of post-investment uncertainty about the capturability of returns. To represent the notion of these costs in this analysis, assume the value per unit of specific human capital in the firm to be $(m+\eta)$, where m is now the expected value of the real productivity, and η is a random component with a density function $\phi(\eta)$ with $E(\eta)=0$. Let ε be the value per unit of specific human capital in the alternative employment, where ε is a random component with a density function $\psi(\varepsilon)$ with $E(\varepsilon)=0$. Realized values of η and ε reflect errors in predicting market conditions facing the firm and the alternative employment, errors in predicting the capacity of the worker to learn new skills on the job, or the influence of real shocks in the economy. To simplify the analysis, I assume that $\mathrm{Cov}(\eta, \varepsilon)=0$, namely that factors affecting the worker's value in the firm are independent of those affecting his alternative value. The employer and the worker are viewed as inferring from past experience the values of m and of the parameters of the density functions. Since our analysis is concerned with fluctuations in the values of specific rather than general human capital, assume the value of the general human capital (H) to remain fixed at unity.

I shall continue to assume that the parties transact costlessly on m and y.[4] Thus, the parties may agree to use some market indicators, sales performance or prevailing wages, for example, in assessing m and y. Fluctuations caused by η and ε are assumed to be known to the respective parties, but are not profitable to negotiate because of high-transaction costs involved in curbing the suspicion that one party is trying to appropriate a portion of the other party's return, or in coping with mutual mistrust in the accuracy of each other's measurements.[5] To simplify the analysis, let us assume that only the employer reacts to the actual value of η and that only the worker reacts to the actual value of ε.

The actual value $\hat{v}$ of the worker to the firm becomes

$$\hat{v}=H+(m+\eta)h=v+\eta h \tag{7}$$

and the actual value $\hat{y}$ to the worker of the alternative employment is given by

$$\hat{y}=H+\varepsilon h \tag{8}$$

Given equations (7) and (8), the jointly optimum separation rule would be given by

$$\hat{v}-\hat{y}\leqslant 0 \text{ i.e., } m\leqslant(\varepsilon-\eta) \tag{9}$$

Note that this rule would apply only if the parties could transact on $\hat{v}$ and $\hat{y}$ costlessly.

The worker will have the incentive to quit when $w-\hat{y}\leqslant 0$, which by substituting (5) and (8) becomes

$$\varepsilon\geqslant\alpha m\equiv\varepsilon^* \tag{10}$$

The employer will have the incentive to dismiss the worker when $\hat{r}\leqslant 0$, which by substituting (5) and (9) becomes

$$\eta\leqslant-(1-\alpha)m\equiv\eta^* \tag{11}$$

[4] Here m and y are assumed to be fixed.

[5] To elaborate this point further, the employer may realize accurately that his business is worse than some agreed-upon indicators show. It may be very costly, however, to convince the worker of the situation and persuade him to accept a wage cut. Alternatively, the worker may perceive that the value of his leisure increased in the second period. If the worker could convince the employer of this increase, the employer might increase his wage to retain the worker. Given the absence of readily available measures of the value of leisure, such a transaction may be too costly to occur.

Thus, the parties may experience separations which would not occur if fluctuations in η and ε could be transacted costlessly. In general the quit and dismissal decisions will be different from each other, and given the underlying assumption of imperfect transactions, the "optimal" decision given by (9) is not attainable.[6] The parties impose external effects on each other by unilaterally separating, and inevitably cause a partial dissipation of the return from the investment. They will choose α to minimize the dissipation of the return.[7] It is worth emphasizing that the basis for the sharing arrangement is the positive costs of transacting on the realized values of η and ε, not the *ex ante* uncertainties about their values.

To determine the optimum sharing ratio, let us consider the objective function. The worker's expected gross gain (M_w) from the investment is given by

$$(12a)\quad M_w=(1-L)(1-Q)E(w)+(1-L)QE(\hat{y}\,|\,\varepsilon>\varepsilon^*)+LE(\hat{y})-H$$

where ε^* is defined by (10), L and Q are probabilities of dismissal and quit, respectively. The gross gain is the sum of values associated with three mutually exclusive and exhaustive outcomes: no dismissal and no quit; no dismissal but quit; dismissal. The worker seeks to maximize his expected net gain given by

$$(12b)\quad G_w=M_w/(1+i)-\beta C\quad 0\leqslant\beta\leqslant 1$$

where i is the interest rate and β is the worker's sharing ratio in the cost of the investment. The employer's gross gain from the investment is the value associated with the outcome of no separation, and is given by

$$(13a)\quad M_e=(1-L)(1-Q)E(\hat{r}\,|\,\eta>\eta^*)$$

where η^* is defined by (11). The employer seeks to maximize his expected net gain given by

$$(13b)\quad G_e=M_e/(1+i)-(1-\beta)C$$

As it is clear by now, the sharing ratio α affects the expected net gain to each party not only by determining the division of the increased productivity between the parties, but also by influencing the quit and the dismissal probabilities. As noted earlier, quits and dismissals produce external effects by preventing the parties from capturing the full return to the investment. The worker is fully aware that the larger the share he claims, the greater will be the probability of dismissal. The employer is just as aware that the larger the share he claims, the greater is the probability of quit. Given *zero* transaction costs in the first period, the worker and the employer choose α and h to maximize the sum of their expected net gains. By maximizing the sum and dividing it between them, they are both better off than if each separately were to maximize his net gain. The sum of the net gains is given by

$$(14)\quad G=M/(1+i)-C\quad M\equiv M_w+M_e$$

Under competition the parties choose α and h by maximizing the sum of their net gains given by equation (14).[8] By maximizing

[6] In this two-period model, all separations are permanent. In reality, cyclical downturns produce both temporary layoffs and permanent separations depending on the decision makers' assessment of the present values of alternative decisions. As Becker (pp. 21–33) discussed, these present values are affected by the severity of the decline in demand, by the expected duration of the decline, and by the prevalence in the economy of the downturn. In my model, the second-period values of $\hat{v}$ and $\hat{y}$ correspond, respectively, to the present values of the productivity in the firm and elsewhere. This model is obviously limited because it is not equipped to analyze temporary separations. However, the main consideration in the sharing arrangement is permanent separations, which result in a complete loss of future returns to the specific capital.

[7] In an interesting recent paper, Mortensen effectively assumed zero transaction costs and arrived at the inevitable conclusion that the quit and the dismissal decisions are independent of the division of specific capital (compare pp. 77–78). See my equivalent discussion earlier and a related discussion in Becker (p. 21).

[8] In the present model, the decision about the number of workers to be employed is ignored. Ignoring this aspect does not crucially affect the analysis of the sharing decision, however. Introducing the decision about the number of workers will complicate the analysis, for example, by making m to be a declining function

G, the parties in effect minimize the dissipation of the return from suboptimal separation.

The first-order equilibrium conditions are given by

$$(15a)\quad \partial G/\partial\alpha=[-L'/(1+i)](1-Q)[\varepsilon^*-E(\varepsilon|\varepsilon<\varepsilon^*)]h+[Q'/(1+i)](1-L)[\eta^*-E(\eta|\eta>\eta^*)]h=0$$

where $L'=\partial L/\partial\alpha>0$ and $Q'=\partial Q/\partial\alpha<0$, and

$$(15b)\quad \partial G/\partial h=[1/(1+i)]\{(1-L)(1-Q)[m+E(\eta|\eta>\eta^*)]+(1-L)QE(\varepsilon|\varepsilon>\varepsilon^*)\}-C'=0$$

The Appendix provides some steps used in deriving these relationships.

The first and the second terms in (15a) are the effects of increased α on the losses, respectively, from a suboptimal dismissal and from a suboptimal quit. This condition may be appreciated by considering the effect of α on the probabilities (L) of a dismissal and (Q) of a quit. An increased α increases L but decreases Q; that is, an increased worker's share in the return to specific human capital increases the probability of a dismissal and decreases the probability of quit. Thus, α is chosen by balancing these opposing effects.[9] Condition (15b) states that the scale of investment is determined by equating the marginal revenue with the marginal cost.

I have shown that the return from the investment will be shared generally. I now discuss the determination of the sharing of the cost of the investment. The sharing of the cost reflects the long-run competitive equilibrium condition requiring that the present value of the return exactly equals costs.[10] In other words, the net gain to the worker (G_w), to the employer (G_e), and therefore to both parties (G), will all equal zero. In particular,

$$(16a)\quad G_w=M_w/(1+i)-\beta C=0$$

$$(16b)\quad G=G_w+G_e=M/(1+i)-C=0$$

$$0\leq\beta\leq 1$$

It follows from (16a) and (16b) that

$$\beta=[M_w/(1+i)]/C=M_w/M$$

or that the worker's share in the cost of investment is equal to the ratio of his return to the total return. Therefore, as long as the return is shared, the cost is shared, and the worker's wage profile is rising. His first-period wage is lower than the alternative wage, and the second-period wage is somewhere between the values of marginal product in this and in alternative firms.

The equilibrium condition given by (15a) shows how the distributions of the productivities in the firm and elsewhere affect the sharing decision. Let us discuss three cases. If both η and ε are degenerate at zero, L, L', Q, and Q' will all equal zero. In this case $\partial G/\partial\alpha$ will be zero always, and α is indeterminate. Put another way, if the employer will not dismiss the worker and the worker will not quit voluntarily, that is, if there is no post-investment uncertainty, the sharing ratio is indeterminate. Indeed, the sharing ratio is not of any economic interest in this case.

Suppose that ε alone is degenerate at zero so that $Q=Q'=0$ always. In this case, the employer may dismiss the worker in the sec-

of the number and by altering the cost function. While the scale of investment (h) will be affected, the sharing decision, which is independent of the scale in my formulation (compare equation (15a)), will not be substantially affected by introducing these complications.

[9] In determining α, the emphasis here, as in most human capital literature, is on the post-investment uncertainty. As Eastman points out, however, uncertainty about the worker's ability to successfully complete the training program will also affect the sharing decision. A worker will impose losses on the employer by failing in the training program, just as he does by quitting in the post-investment period. The employer has the additional incentive, therefore, to shift to the worker the investment costs and returns.

[10] All firms are assumed to be identical with respect to the cost function for investment and other relevant aspects, even though the human capital, once invested, is specific to each firm. Therefore, the equilibrium condition at the market level can be inferred from the zero profit condition for a single firm.

ond period, but the worker will not quit. Equation (15a) then becomes

$$(15c)\qquad -L'(\alpha m)=0$$

which states that the parties minimize the loss from dismissal. Clearly, the optimum sharing arrangement will set $\alpha=0$. Knowing that he may be dismissed, the worker has no incentive to share the investment. His wage profile is flat, and he receives the alternative wage (y) in both periods.

Consider the other extreme case in which η alone is degenerate at zero so that $L=L'=0$ always. In this case, the employer will not dismiss the worker, but the worker may quit. Equation (15a) now becomes

$$(15d)\qquad -Q'(1-\alpha)m=0$$

which states that the parties minimize the loss from quit. The optimum sharing arrangement will set $\alpha=1$. Thus, the worker receives all of the return, and by implication, pays all of the cost. Knowing that the worker may quit, the employer has no incentive to share in the investment. The worker's wage profile is rising, reflecting the full value of his marginal product in this firm.

In general, neither ε nor η is degenerate at zero, and both the employer and the worker potentially have the incentive to separate. As a result, the investment is shared, and the worker's wage profile is rising but is below his value of marginal product in this firm during the post-investment period.

II. Concluding Remarks

Firm-specific human capital is shared when both the worker and the employer potentially have the incentive to separate even though both together will be made worse off by the separation. This paper discusses a two-period model in which the parties attempt to minimize the loss from such separations by optimally sharing the investment. The analysis reveals that whether or not the investment is shared depends on the existence of transaction costs in evaluating and agreeing on the worker's productivities. This conclusion is a straightforward application of the Coase Theorem to the income streams generated by specific human capital.

While the model presented here extends the original Becker analysis and the subsequent development in the literature, further extensions and elaborations, beyond the scope of this paper, are worth noting. Let me mention three examples. First, the model could be expanded to include more than two periods. A multiperiod model would facilitate an analysis of both temporary and permanent separations as they are affected by specific human capital (compare fn. 6). Second, one may wish to cast the analysis in a general equilibrium framework in which the decisions about the number of workers (compare fn. 8), amounts of physical capital, for example, play more explicit roles than they do in the present model. Finally, the present analysis represents the influence of transaction costs simply as random fluctuations in the values of productivities. While this approach simplifies the analysis, a detailed examination of the nature of transaction costs involved in the employment contract will increase our understanding of wage profiles, labor turnover, and other labor market phenomena.[11]

Appendix

The purpose of this Appendix is to indicate some crucial steps in deriving the first-order conditions.

Given that

$$\hat{v}=H+(m+\eta)h$$

$$\hat{y}=H+\varepsilon h$$

$$\text{(A1)}\quad E(\hat{v}\mid\eta<\eta^*)=H+mh+h\frac{\int_{min}^{\eta^*}\eta\phi(\eta)d\eta}{\int_{min}^{\eta^*}\phi(\eta)d\eta}$$

$$=H+mh+\frac{NUM}{L}h$$

$$NUM\equiv\int_{min}^{\eta^*}\eta\phi(\eta)d\eta,\ L\equiv\int_{min}^{\eta^*}\phi(\eta)d\eta$$

$$\therefore\quad\frac{\partial E(\hat{v}\mid\eta<\eta^*)}{\partial\alpha}$$

[11] See my article with Yu for related discussions.

$$=h\left\{\frac{L\eta^*\phi(\eta^*)-\phi(\eta^*)NUM}{L^2}\right\}\frac{\partial\eta^*}{\partial\alpha}$$

$$=\frac{h\phi(\eta^*)}{L}\{\eta^*-E(\eta|\eta<\eta^*)\}\frac{\partial\eta^*}{\partial\alpha}$$

$\because$ $\partial L/\partial\eta$ evaluated at η^* equals $\phi(\eta^*)$.

By applying similar procedure, one obtains

$$\text{(A2)}\quad \frac{\partial E(\hat{v}|\eta>\eta^*)}{\partial\alpha}=\frac{-h\phi(\eta^*)}{1-L}\times\{\eta^*-E(\eta|\eta>\eta^*)\}\frac{\partial\eta^*}{\partial\alpha}$$

$$\text{(A3)}\quad \frac{\partial E(\hat{y}|\varepsilon<\varepsilon^*)}{\partial\alpha}=\frac{h\psi(\varepsilon^*)}{1-Q}\times\{\varepsilon^*-E(\varepsilon|\varepsilon<\varepsilon^*)\}\frac{\partial\varepsilon^*}{\partial\alpha}$$

$$\text{(A4)}\quad \frac{\partial E(\hat{y}|\varepsilon>\varepsilon^*)}{\partial\alpha}=\frac{-h\psi(\varepsilon^*)}{Q}\times\{\varepsilon^*-E(\varepsilon|\varepsilon>\varepsilon^*)\}\frac{\partial\varepsilon^*}{\partial\alpha}$$

Also recall that

$$\text{(A5)}\quad E(\hat{v})=LE(\hat{v}|\eta<\eta^*)+(1-L)E(\hat{v}|\eta>\eta^*)$$
$$E(\hat{y})=QE(\hat{y}|\varepsilon>\varepsilon^*)+(1-Q)E(\hat{y}|\varepsilon<\varepsilon^*)$$

(To derive (15a) and (15b).)

$$\text{(A6)}\quad \frac{\partial G}{\partial\alpha}=\frac{1}{1+i}\frac{\partial M}{\partial\alpha}=0$$

but
$$\frac{\partial M}{\partial\alpha}=-\frac{\partial L}{\partial\alpha}(1-Q)E(\hat{v}|\eta>\eta^*)-(1-L)\frac{\partial Q}{\partial\alpha}E(\hat{v}|\eta>\eta^*)+(1-L)(1-Q)\frac{\partial E(\hat{v}|\eta>\eta^*)}{\partial\alpha}-\frac{\partial L}{\partial\alpha}QE(\hat{y}|\varepsilon>\varepsilon^*)+(1-L)\frac{\partial Q}{\partial\alpha}E(\hat{y}|\varepsilon>\varepsilon^*)+(1-L)Q\frac{\partial E(\hat{y}|\varepsilon>\varepsilon^*)}{\partial\alpha}+\frac{\partial L}{\partial\alpha}E(\hat{y})$$

and using the relationships (A1) through (A5), one obtains (15a).

$$\text{(A7)}\quad \frac{\partial G}{\partial h}=\frac{1}{1+i}\frac{\partial M}{\partial h}-C'=0$$

but
$$\frac{\partial M}{\partial h}=(1-L)(1-Q)\frac{\partial E(\hat{v}|\eta>\eta^*)}{\eta h}+(1-L)Q\frac{\partial E(\hat{y}|\varepsilon>\varepsilon^*)}{\partial h}+L\frac{\partial E(\hat{y})}{\partial h}$$
$$=(1-L)(1-Q)[m+E(\eta|\eta>\eta^*)]+(1-L)QE(\varepsilon|\varepsilon>\varepsilon^*)$$

from which (15b) is directly obtained.

REFERENCES

G. S. Becker, "Investment in Human Capital: A Theoretical Analysis," *J. Polit. Econ.*, Oct. 1962, *70*, 9–49.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

D. Donaldson and B. C. Eaton, "Firm-Specific Human Capital: A Shared Investment or Optimal Entrapment?," *Can. J. Econ.*, August 1976, *9*, 462–72.

S. Eastman, "Uncertainty and the Time Profile of Wages," *Can. J. Econ.*, Aug. 1977, *10*, 472–73.

M. Hashimoto, "Bonus Payments, On-The-Job Training, and Life-time Employment in Japan," *J. Polit. Econ.*, Oct. 1979, *87*, 1086–104.

_______ **and B. T. Yu**, "Specific Capital, Employment Contracts and Wage Rigidity," *Bell J. Econ.*, Autumn 1980, *2*, 536–49.

D. T. Mortensen, "Specific Capital and Labor Turnover," *Bell J. Econ.*, Autumn 1978, *9*, 572–86.

D. O. Parsons, "Specific Human Capital: An Application to Quit Rates and Layoff Rates," *J. Polit. Econ.*, Nov./Dec., 1972, *80*, 1120–143.

______, "Models of Labor Market Turnover: A Theoretical and Empirical Survey," *Res. Labor Econ.*, 1977, *1*, 185–223.

J. H. Pencavel, "Wages, Specific Training and Labor Turnover in U.S. Manufacturing Industries," *Int. Econ. Rev.*, Feb. 1972, *13*, 53–64.

Fixed Wages, Layoffs, Unemployment Compensation, and Welfare: Note

By HERSCHEL I. GROSSMAN AND KENNETH HAPPY*

In a recent paper, H. M. Polemarchakis and L. Weiss (hereafter P-W) argue that the existence of implicit contractual arrangements for shifting risk from workers to employers causes inefficiency in the allocation of labor resources. Their analysis assumes that there is interindustry variability in demand, that changing jobs is costly, and that employers behave in a risk-neutral manner, absorbing the risk associated with disturbances to the pattern of demand and providing risk-averse workers with a constant net income. According to P-W, under conditions of competitive equilibrium, such risk-shifting arrangements would involve two practices: First, workers would receive a constant wage rate equal to the average value of their marginal product. Second, employers would pay the job-changing costs of any workers whom they laid off when they faced depressed demand.

The problem alleged by P-W is that this second practice would unduly restrict labor mobility, because it would induce employers to keep employment too high when demand in their industry was depressed. Specifically, employers would maintain employment under low demand such that the value of labor's marginal product under low demand plus the amount of job changing costs was equal to the average value of marginal product rather than the value of marginal product under high demand. This argument leads P-W to conclude that a partial public subsidy of job changing would improve resource allocation.

The P-W results are not consistent with Robert Barro's argument that, in competitive equilibrium, contractual arrangements for shifting risk are consistent with efficient resource allocation. P-W attribute the inefficiency that they associate with risk-shifting arrangements to job-changing costs, which Barro does not explicitly consider, and the essential element in their analysis is the contention that in competitive equilibrium employers would provide severance pay equal to the job-changing costs of anyone they laid off.

The purpose of this note is to point out that P-W's results require the implicit, and apparently unrealistic, assumption that employers in expanding industries cannot offer to compensate laid-off workers for part of their job-changing costs. We argue that, without this implicit assumption, the existence of job-changing costs would not undermine Barro's argument. We derive a competitive equilibrium in which employers in contracting and expanding industries are sharing the job-changing costs, labor mobility is not unduly restricted, and resource allocation is efficient.

The formal analysis in the P-W paper involves a model of two industries and two periods. In the first period, labor is allocated such that the product of output price and marginal product is equal to the wage rate and is the same in both industries. In the second period, a symmetrical shift in the pattern of demand will make output price higher in one industry and lower in the other industry, but the direction of the demand shift is random. Workers can move from the low-price industry to the high-price industry in response to this demand shift, but each job change involves a positive cost, given by c.

Employers absorb all of the risk associated with the randomness in the pattern of demand. According to P-W, this risk shifting requires a constant wage rate and severance pay equal to c. In the second period, these arrangements induce employers in the industry that faces a low output price to lay off workers only until their value of marginal

*Brown University. The National Science Foundation has supported the research on which this comment is based.

product equals the wage rate minus c. Given the symmetry of the demand shift, this limited release of labor to move to the industry that faces high output price would leave the value of marginal product in this expanding industry equal to the wage rate plus c. The resulting interindustry difference in value of marginal product is $2c$, whereas a job change costs only c. Thus, an efficient outcome would require further reallocation of labor.

Polemarchakis and Weiss infer from this analysis that, with job-changing costs, risk shifting is inconsistent with efficient resource allocation. Their analysis, however, neglects the full implications of the fact that with the outcome they propose profit-seeking employers in the expanding industry would want to bid for additional workers. They rule out one form that this bidding might take by explicitly assuming that employers cannot pay higher wages to new workers than to old workers. Their explicit assumptions, however, do not rule out other recruiting efforts that involve lump sum payments in cash or kind to new workers. Common examples of such payments include reimbursement for job-changing costs such as moving expenses and provision of services such as help-wanted advertisements that mitigate job-changing costs. The P-W analysis does not recognize the relevance of these standard recruiting practices and implicitly precludes them.

The important effect of these recruiting practices is that they reduce the amount of severance pay that is required in order to relieve workers of the risk of demand disturbances. In the specific model considered by P-W, employers in the expanding industry would have an incentive to increase recruiting efforts until they are in effect making payments of $c/2$ to each laid-off worker. Employers in the contracting industry, consequently, would only have to provide severance pay equal to the same amount. This outcome would induce the quantity of layoffs and job changes that would raise the value of marginal product in the contracting industry to the wage rate minus $c/2$ and lower the value of marginal product in the expanding industry to the wage rate plus $c/2$. The resulting interindustry difference in value of marginal product would be equal to c, the cost of an additional job change.

In this equilibrium, employers still absorb all risk, but the possibilities for profitable recruiting efforts are exhausted and, as a result, the expected value of marginal product, expected worker income, and expected profits are all higher than in the outcome proposed by P-W. Employers in the expanding and contracting industries are each covering an appropriate share of the job-changing costs of laid-off workers, layoffs, and labor mobility reflect the actual costs of job changing, and resource allocation is efficient. Risk shifting and job-changing costs do not provide a rationale for a public subsidy of job changing.

REFERENCES

R. Barro, "Long-Term Contracting, Sticky Prices, and Monetary Policy," *J. Monet. Econ.*, July 1977, *3*, 305–16.

H. Polemarchakis and L. Weiss, "Fixed Wages, Layoffs, Unemployment Compensation, and Welfare," *Amer. Econ. Rev.*, Dec. 1978, *68*, 909–17.

The Technology of Risk and Return: Comment

By CHRISTOPHER JAMES*

In a recent article in this *Review*, Edward Greenberg, William Marshall, and Jess Yawitz (hereafter, G-M-Y) analyze the relationship between the firm's behavior in the product market and capital market conditions. Using the mean-variance or capital asset pricing model (*CAPM*), G-M-Y argue that the firm can be viewed as operating along a concave mean-covariance efficiency frontier. They contend that product market decisions will affect the mean and covariance of the firm's dollar returns with the market portfolio consisting of all risky assets. They further argue that value-maximizing firms will choose to operate where the marginal tradeoff between mean and covariance along the efficiency frontier is equal to the market price of risk. They proceed to illustrate the usefulness of their analysis by examining a number of product market decisions including diversification, capital budgeting, pricing, and output decisions.

The purpose of this paper is to show that the assumptions needed to derive G-M-Y's results are *not* consistent with the *CAPM*. Three major criticisms are made of the G-M-Y analysis:

1) Firms are assumed to take the dollar price of risk as given while recognizing that new investment or other product market changes will impact upon the future return distribution of the market portfolio. This form of analysis makes sense if and only if investors' utility function exhibit constant absolute risk aversion. Hence they are dealing with a special case of the *CAPM*. In general, the price of risk cannot be assumed exogenous given a change in the market portfolio.

2) Concavity of the mean-covariance efficiency frontier can be established if and only if the firm is *not* perceived a price taker in the capital markets. Unfortunately, if imperfect competition is assumed the *CAPM* is no longer an appropriate tool for analyzing capital market equilibrium. By assuming imperfectly competitive capital markets, decisions which maximize firm value cannot be unambiguously defined unless homogeneous preferences of investors is assumed. In other words, with imperfectly competitive markets the effect of the firm's production and investment decisions on the value of its shares cannot be determined without reference to investors' wealth and preferences.[1]

3) Even assuming constant absolute risk aversion the G-M-Y analysis implicitly assumes that other firms in the market are in disequilibrium.

The first point can be established by recognizing that the *CAPM* describes the equilibrium value or return on assets when capital markets are assumed perfect and the quantities of assets are fixed.[2] The equilibrium value of the firm can be written as

$$(1)\quad S_i = \frac{1}{R_f}\left[E(\tilde{x}_i) - \lambda_m \, cov(\tilde{x}_i, \tilde{x}_m)\right]$$

where

S_i = present value of firm i at beginning of period 1
R_f = risk-free interest factor
$\tilde{x}_i$ = dollar returns of firm i at the end of period 1
λ_m = market price of risk
cov = covariance operator
$\tilde{x}_m$ = dollar return on the market portfolio at the end of period 1.

*Assistant professor of finance and business economics, University of Oregon. I would like to thank M. Hopewell, L. Dann, G. Racette, and G. O. Bierwag for their assistance.

[1] David Baron, in a article published after this paper was written, provides an excellent review of the literature on the optimality of the investment allocation resulting from firm value maximization within the context of the mean-variance model.

[2] See Michael Jensen, N. C. Nielsen, or Eugene Fama (1972) for a derivation of the *CAPM* as well as a list of assumptions used in deriving the *CAPM*.

The market price of risk in equilibrium can be expressed as

$$(2) \qquad \lambda_m = \left(\sum_{k=1}^{N} \phi_k \right)^{-1}$$

where ϕ_k = the marginal rate of substitution between expected return and variance for the kth investor. Or as Mark Rubinstein (1973a) has shown,

$$(3) \qquad \lambda_m = \left(\sum_{k=1}^{N} \theta_k \right)^{-1}$$

$$\text{where} \qquad \theta_k = \frac{E[U_k''(\tilde{\omega}_k)]}{E[U_k'(\tilde{\omega}_k)]}$$

and $\tilde{\omega}_k$ = end of period wealth of investor k; U= utility function of investor k.

Clearly, as equations (2) and (3) illustrate, the equilibrium price of risk is a function of both investors' preferences and the aggregate level of wealth. Therefore, in analyzing decisions which impact investors' aggregate wealth or the distribution of returns on the market, the market price of risk will be constant if and only if investors are assumed to have utility functions which exhibit constant absolute risk aversion.[3] G-M-Y assume that the market price of risk is constant even as the composition of the market portfolio changes. They therefore limit their analysis to a special case of the *CAPM* by implicitly assuming constant absolute risk aversion.

If the quantities of risky assets are fixed and the distribution of the market portfolio is given, then the dollar price of risk can also be written as

$$(4) \qquad \lambda_m = \frac{E(\tilde{x}_m) - R_f S_m}{\sigma_m^2}$$

where σ_m^2 = variance of dollar returns on the market portfolio; S_m = value of the market portfolio.

But, equation (4) may be used to express the market price of risk if and only if perfect competition in the capital market is assumed. Perfect competition is defined to exist if each firm acts as if *individually* its marginal supply does not affect the risk-free rate of interest, the market price of risk or its covariance with the market *through changes in aggregate output*. This definition is consistent with the one used recently by Robert Merton and Marti Subrahmanyam in which the firm is assumed a price taker if it perceives a horizontal supply curve for capital. Both definitions are equivalent to assuming the firm cannot affect the aggregate amount of investment undertaken by all firms or the distribution of returns on the market portfolio. Perfect competition is therefore equivalent to assuming the quantity of risky assets and the distribution of returns on the market portfolio are given. Only under these conditions can the price of risk be expressed (as in equation (4)) without reference to individual investor's preferences. But G-M-Y's analysis allows both the quantity of risky assets and the composition of the market portfolio to change as a result of the decisions of the firm under scrutiny, and thus they violate the perfect capital market assumption on which the *CAPM* is based.

The significance of this violation for their analysis is difficult to overstate, since I will now demonstrate that the alleged concavity of the mean-covariance efficiency frontier stems from the implicit assumption of imperfect competition and not from any technological relationship as G-M-Y imply. I demonstrate that if perfect competition, as defined above, is assumed then the efficiency frontier is linear in mean-covariance space. Not surprisingly, this result is consistent with the conclusions of Rubinstein (1973b) and others using the *CAPM*. To establish this result, I will analyze two issues addressed by G-M-Y: investment under constant stochastic returns to scale and firm diversification.

The assumption that capital markets are competitive and that firms cannot affect the quantities of risky assets permits presentation of the equilibrium value of the firm in a slightly different form than that of G-M-Y: the equilibrium expected return of the jth

[3] See John Lintner for a discussion of this point.

firm is

$$(5)\qquad E(\tilde{R}_j)=\frac{E(\tilde{x}_j)}{S_j}=R_f+\lambda^* cov(\tilde{R}_j,\tilde{R}_m)$$

where $E(\tilde{R}_j)$=expected return of asset j; $E(\tilde{x}_j)$=expected dollar return of firm j;

$$\lambda^*=\frac{E(\tilde{R}_m)-R_f}{(\sigma_m^2)^*}$$

and $(\sigma_m^2)^*$=variance of returns on the market portfolio.

Equation (5) can be used to solve for the value of the firm by noting that $\lambda^*=S_m\lambda_m$. Since perfect competition is assumed, S_m is constant as the firm alters its product market decisions. The value of λ^* will therefore be constant given perfect competition and the value of the firm may be written as

$$(6)\quad S_j=\frac{1}{R_f}\left\{E(\tilde{x}_j)-\lambda^* cov(\tilde{x}_j,\tilde{R}_m)\right\}$$

Equation (6) can be used to evaluate the firm's investment and diversification decisions. First the firm's investment decisions under constant stochastic returns to scale will be evaluated. If ρ equals the stochastic return per dollar invested and I equals the dollar amount invested in the project, then the value of the firm upon undertaking the project may be written as[4,5]

$$S_j=\frac{1}{R_f}\left\{E(\tilde{x}_j)+E(\rho)I-\lambda^* cov(\tilde{x}_j+I\rho,\tilde{R}_m)\right\}-I$$

and

$$(7)\quad S_j=\frac{1}{R_f}\left\{E(\tilde{x}_j)+E(\rho)I-\lambda^* cov(\tilde{x}_j,R_m)-\lambda^* I\, cov(\rho,\tilde{R}_m)-R_f I\right\}$$

Equation (7) reflects the assumption that the value of the market portfolio, S_m, as well as the return on the market portfolio is constant. Equation (7) also indicates that the optimum level of investment in the project *for the firm* is indeterminate. By differentiating equation (7) with respect to I and setting the result equal to zero, we obtain

$$(8)\qquad E(\rho)=R_f-\lambda^* cov(\rho,R_m)$$

Note that the optimum level of investment in the project is independent of the return of the firm's existing assets. This result should *not* be surprising, since as Stewart Myers, Robert Hamada, Rubinstein (1973b), and others have shown, in perfect capital markets for nonsynergistic investment projects the acceptance criteria is not a function of the returns on existing assets.

In addition, it can be seen that the volume of investment by a single firm in a competitive capital market is indeterminate for the case of constant stochastic returns to scale.

[4] Equation (7) differs significantly from the expression for the value of the firm assuming imperfect competition, G-M-Y express the value of the firm upon acceptance of the project as

$$(a)\qquad S_j=\frac{1}{R_f}\left\{E(\tilde{x}_j)+(E(\rho)-R_f)I-\lambda_m(\sigma_{jm}+I\sigma_{j\rho}+I\sigma_{\rho m}+I^2\sigma_\rho^2)\right\}$$

which clearly demonstrates that the firm is perceived to have an impact on the market portfolio.

[5] Merton and Subrahmanyan obtain the same result through a slightly different procedure which illustrates the effect of holding aggregate investment constant. Let I_j equal the amount of investment in the project undertaken by firm j and let I equal the total amount of investment. The value of the firm can be written as

$$(b)\qquad S_j(I_j,I)=S(O,I)+I_j g(I)$$

where

$$g(I)=\frac{1}{R_f}\left[E(\rho)-\lambda(\sigma_{m\rho}+I\sigma_\rho^2)\right]$$

$$S_j(O,I)=\frac{1}{R_f}\left[\tilde{x}_j-\lambda(\sigma_{mj}+I\sigma_{j\rho})\right]$$

$S_j(O,I)$ represents market value of assets of firm j given investment level I, and $g(I)$ is value per unit of investment in the new project. Differentiating (b) with respect to I_j indicates that if $dI/dI_j=0$ (total investment is independent of the action of firm j) then the value-maximizing level of I_j is indeterminate. If however, the firm can affect aggregate investment, then the investment project is not independent of the firm's assets because changes in aggregate investment alters the cost of capital associated with the firm's existing projects.

This result is analogous to the well-known result for competitive product markets with constant return to scale production technology: the output of a single firm is indeterminate while aggregate industry output can be determined. Note that equation (7) also indicates that the mean-covariance efficiency frontier for a competitive firm is linear. The slope of the efficiency frontier, as expected, is equal to that of the security market line.

Similar criticisms can be made of G-M-Y's analysis of diversification, that is, that the firm is assumed to be an imperfect competitor in the capital market in order to derive the concave efficiency frontier. G-M-Y consider the case in which a firm operates simultaneously in two industries and is contemplating expansion in industry one at a rate of Ψ_1, and in industry two by Ψ_2, G-M-Y obtain the following expression for the firm's mean and covariance of dollar returns:

$$(9)\qquad \mu=(1+\Psi_1)\mu_1+(1+\Psi_2)\mu_2$$

$$(10)\quad cov=(1+\Psi_1)\sigma_{1m}+(1+\Psi_2)\sigma_{2m} +\Psi_1(1+\Psi_1)\sigma_{11}+\Psi_2(1+\Psi_2)\sigma_{22} +(\Psi_1+\Psi_2+2\Psi_1\Psi_2)\sigma_{12}$$

Equation (10) indicates that the impact of the firm's investment on the market is being considered in the expansion decision. If perfect competition is assumed, then the firm's covariance with the market can be written as

$$(11)\quad cov=(1+\Psi_1)\sigma_{1m}+(1+\Psi_2)\sigma_{2m}$$

Obviously no incentive exists for diversification, since the covariance between industries does not enter into the firm's decision. This result is consistent with the conclusions of other authors using the *CAPM* (or other valuation models); in perfect capital markets, since investors are able to diversify for themselves, firms will not be rewarded for undertaking diversification.[6] While G-M-Y attribute their results to a technological externality in the capital market, it is apparent that their result is due to assuming imperfect competition. Again, under perfect competition, the efficiency frontier is perceived by the firm as linear, and the firm will make decisions according to this linear frontier.

So far I have shown that if a firm *perceives* its actions will not affect the covariance of its returns with the market through changes in aggregate supply, then the firm operates along the security market line. Since perceived price independence is an assumption, one need not ask the conditions necessary for its occurrence. If, alternatively, the analysis is pursued based upon *actual* price independence, the question must be raised as to the conditions necessary for the firm to have no impact on the aggregate amount of risk and return. The literature in this area concludes that either other firms in the market must react in such a way as to exactly offset the impact of the firm or aggregate output or, alternatively, the number of firms is so large that in the limit the impact of the firm's decision approaches zero.[7]

The question can be raised, however, as to the effect of imperfect competition on the structure of equilibrium capital market prices and whether with imperfect competition the *CAPM* is an appropriate tool to analyze firm behavior. In general it is not, since with imperfect competition the Fisherian separation of production and consumption decision no longer holds. In such a world the firm cannot, without information as to stockholder preferences and the distribution of wealth, determine whether its actions maximize the value of its shares.

To see this, recall that the market price of risk is equal to the inverse of the summed marginal rates or substitution between expected return and variance over all investors or alternatively is equal to the summed Pratt-

[6] See Fama and Merton Miller for a discussion of this point.

[7] See Fama (1972), Rubinstein (1978) and Baron. The "reaction" specified by Fama may not, however, be consistent with actual price independence, since the firm may still alter the distribution of returns on the market portfolio. See Nielsen for a discussion of this point. For the conditions necessary for actual price independence within the context of a state preference model, see Harry DeAngelo.

Arrow measure of risk aversion. If the firm alters the aggregate amount of return and risk in the market, this in general will affect investors end of period wealth. This in turn will affect the market price of dollar risk which will change the present value of the firm. The impact of the firm's behavior on the price of risk will depend upon investors' preferences and wealth as equation (2) indicates.[8] In short, if the firms possess market power as described above, then the effect of the firm's investment decision on the present value of the firm cannot be determined without reference to investors' preferences.

It may appear that, since their analysis must assume constant absolute risk aversion by all investors, G-M-Y may avoid the ambiguities discussed above. Unfortunately, even under this set of assumptions, the analysis presented is not consistent with capital market equilibrium. To the extent that the firm alters the risks and returns on the market portfolio of assets, this will imply that all other firms are not operating along the security market line. Therefore, to the extent that other firms alter their decisions, this implies reoptimization on the part of the original firm. By making a Cournot-type assumption about the behavior of other firms, G-M-Y are assuming the economy is *not* operating in equilibrium.[9] In defense of G-M-Y, one might argue at this point that ignoring disequilibrium is a standard partial equilibrium *ceteris paribus* assumption, frequently employed in theory of the firm analysis. However, as I have demonstrated, G-M-Y incorrectly use a general equilibrium model (the *CAPM*) for their partial equilibrium analysis.

I have shown that the analysis of G-M-Y is not consistent with the *CAPM* and their conclusions follow only if the inconsistencies are maintained. If firms can affect the aggregate amount of risk and return in the market, then without detailed information about investors' preferences, the firm cannot determine the impact of its behavior on the present value of its shares. One of the major appeals of the *CAPM* is that it describes the equilibrium structure of prices without using investors' preferences. However, this does not imply that the determination of equilibrium prices are independent of investors' risk preferences.[10] Further, if firms possess monopoly power in the capital market, then the analysis should include not only the impact upon aggregate supply, but also evaluate the firm's decision under a different set of relative prices consistent with the new aggregate supply. These relative prices cannot, however, be determined without reference to investors' preferences.

[8] If the firm can affect the aggregate risk and return in the market, then this may also affect the risk-free rate of interest since in equilibrium

$$\frac{\partial U}{\partial C_k} \Big/ \frac{\partial U}{\partial E(\tilde{w}_k)} = R_{f\forall k}$$

where C_k = present consumption by the kth investor. Since the firm may effect $E(\tilde{w}_k)$, an alternative risk-free rate is implied.

[9] Equation (b) in fn. 4 can be used to demonstrate that changes in aggregate investment will affect the cost of capital of other firms in the market and lead to a change in investment.

[10] For a description of the process by which equilibrium prices are established, see Fischer Black.

REFERENCES

D. P. Baron, "Investment Policy, Optimality and the Mean Variance Model," *J. Finance*, Mar. 1979, *34*, 207–32.

F. Black, "Equilibrium in the Creation of Investment Goods Under Uncertainty," in Michael Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1972.

H. DeAngelo, "Three Essays in Financial Economics," unpublished doctoral dissertation, Univ. California-Los Angeles, 1977.

E. F. Fama, "Perfect Competition and Optimal Production Decisions Under Uncertainty," *Bell J. Econ.*, Autumn 1972, *3*, 509–30.

——, "Risk, Return, and Equilibrium," *J. Polit. Econ.*, Jan./Feb. 1971, *78*, 30–55.

—— **and Merton Miller,** *The Theory of Finance*, New York 1972.

E. Greenberg, W. J. Marshall, and J. B. Yawtiz, "The Technology of Risk and Return," *Amer. Econ. Rev.*, June 1978, *68*, 241–51.

R. S. Hamada, "Portfolio Analysis, Market

The Technology of Risk and Return: Reply

By Edward Greenberg, William J. Marshall, and Jess B. Yawitz*

Christopher James's comment gives us the opportunity to correct an error in our paper: we incorrectly stated that Hayne Leland's (references in the paper) use of the expected utility approach yields results that are dependent on individual preferences.

The main point of James's comment is his opinion that "...G-M-Y incorrectly use a general equilibrium model (the *CAPM*) for their partial equilibrium analysis" (p. 489). We presume that this ambiguous phrase does not mean that in principle it is incorrect to use a general equilibrium model as the starting point for partial analysis, especially since this approach is probably the single most useful tool in applied economics. Therefore, we interpret James's statement to mean that our use of the *CAPM* for this purpose is incorrect. However, James has proved no such thing—he has merely illustrated the principle that different assumptions lead to different conclusions. There is no one correct way to do partial equilibrium analysis, and we believe the way we have chosen is more useful than James'. It is, moreover, the approach taken by a number of other writers in the finance literature.

To be more specific, in our introductory comments and in each of the specific cases we examine, the firm takes as given the risk-free interest rate (R_f), the market price of risk (λ_m), the covariances between its cash flows and those of all other firms, and the decisions of other firms. The firm then selects an expected return-covariance combination ($\mu-\sigma$) by choosing values for those variables under its control.

In the two cases discussed by James, we assume that the firm adds its new output to the market portfolio when determining optimum investment. James states that this procedure violates the perfect capital market assumption of the *CAPM*, but he seems to confuse the nature of competition in the product markets on the one hand and in the capital market on the other. Although James' assumption that the value of the market portfolio is fixed makes sense if product markets are perfectly competitive and at a long-run equilibrium, competition in the capital market is compatible with less-restrictive conditions. Analysis of the investment decision in long-run equilibrium is of little interest, since there are no incentives to invest. Nevertheless James appears to do such an analysis, invoking a rather peculiar specification of the competitive firm. From his equation (7) and the discussion directed to it, it appears that James' firm acts as if it knows its decision will leave the value of the market portfolio unchanged, but this can occur only if an increment in its output is exactly offset by declines in that of others. The usual assumption is that each firm ignores the effects of its decisions on its competitors. (Indeed, the implication of n identical firms acting as James portrays them is chaos.) A more detailed discussion of this and related points may be found in Yawitz. Exactly the same problem arises in James's discussion of the diversification problem—he again assumes that other firms offset the actions of the firm being studied.

The approach we take is the same as that taken by Michael Jensen and John Long; the term in I^2 contained in their equation (10) and our equation (7) yields a determinate result for the case of constant stochastic returns to scale. Moreover, the review article by David Baron, cited by James, characterizes the value-maximizing investment exactly as we do, assuming that R_f and λ_m are held constant, but that the firm's new output is included in the market portfolio.

To conclude, we believe that our use of the *CAPM* for partial equilibrium analysis is consistent with current practice. Whether it is worthwhile will depend on the theoretical

*Professor of economics, associate professor of finance, and professor of finance and business economics, respectively, Washington University.

and empirical research it stimulates. It seems to us that the main contribution of the use of the *CAPM* and other tools from modern finance theory is the finding that, under uncertainty, firms' valuations are interdependent. Recognition of this interdependence in the analysis of firm decision making, with the aid of the *CAPM* and more general models, may yield interesting and fruitful hypotheses for further research.

REFERENCES

D. P. Baron, "Investment Policy, Optimality, and the Mean-Variance Model," *J. Finance*, Mar. 1979, *34*, 207–32.

E. Greenberg, W. J. Marshall, and J. B. Yawitz, "The Technology of Risk and Return," *Amer. Econ. Rev.*, June 1978, *68*, 241–51.

C. James, "The Technology of Risk and Return: Comment," *Amer. Econ. Rev.*, June 1981, *71*, 485–90.

M. Jensen and J. Long, "Corporate Investments under Uncertainty and Pareto Optimality in the Capital Market," *Bell J. Econ.*, Spring 1972, *3*, 151–74.

J. B. Yawitz, "Externalities and Risky Investments," *J. Finance*, Sept. 1977, *32*, 1143–49.

Ownership Arrangements and Congestion-Prone Facilities

By DAVID E. MILLS*

This paper concerns a class of facilities or resources whose services to users decrease in value with an increase in the number of users. They will be called congestion-prone facilities. Examples abound in the literature of economics; roads, parks, museums, streams, lakes, hunting preserves, wilderness areas, and common oil and gas pools are but a sample. The effect of congestion on the quality of facility services varies among examples: for roads it increases both the time required to make a trip and the risk of an accident in route, for museums or wilderness areas it diminishes an aesthetic aspect of the service offered, and for common oil and gas pools it decreases rates of return on invested capital. These congestion effects are distinguished from other kinds of externality problems in that external effects are confined to facility users; those who, regardless of congestion considerations, would not use the facility are not affected by its externality problem.

The focus of the paper is a comparison of the efficiency characteristics of two facility ownership arrangements—private and common ownership. In general, neither regime achieves optimal congestion apart from social intervention. That common ownership leads to over-congestion was clearly shown by H. Scott Gordon, although the result has antecedents in A. C. Pigou's discussion of increasing cost industries. The effect of private ownership is more complicated.

The possibility of optimal congestion under private ownership was first shown by Frank Knight. His argument was anecdotal and employed Pigou's example of the two roads—one narrow and fast, the other broad but slow.[1] He showed that optimal congestion on the narrow, congestion-prone road would be achieved by merely conferring ownership rights to it on a single agent, thus enabling him to charge a revenue-maximizing toll.

It has long been known that Knight's claim for the generality of this result was excessive.[2] James Buchanan showed that it held in the road example only because there was an alternative route to the congestion-prone road, and that ownership rights could be conferred without simultaneously, and inadvertently, conferring monopoly power. After describing some alternative institutional arrangements where the result fails to hold, he concluded that private ownership can be relied upon to achieve efficient resource use "[o]nly in those cases where the extent of commonality of usage is limited to a relatively small proportion of the total resource supply..." (p. 315). That is, there must be a sufficient supply of alternatives to the facility in question to prevent monopoly power from existing if it is owned privately. Later, working with a variation of the road example, Noel Edelson qualified Knight's claim in another way.[3] He showed that it depended on all users of the two roads having the same imputed value of time spent in transit.

*Assistant professor of economics, University of Virginia. I would like to thank George Borts, James Buchanan, Roger Sherman, the late Michael Visscher, John Whitaker, and an anonymous referee for helpful comments on earlier versions of this paper.

[1] First introduced by Pigou in *Wealth and Welfare* (1912), this example also appeared in the first edition of the more durable *The Economics of Welfare* (1920). It was dropped from the second edition in 1924, the same year Knight's critique appeared, but has nevertheless enjoyed a long and prosperous life. James Buchanan and Noel Edelson have each contributed substantially to its elucidation.

[2] It is ironic that Knight generalized so freely with this example, having taken Pigou to task for the same error: "Professor Pigou's logic in regard to the roads is, as logic, quite unexceptionable. Its weakness is one frequently met with in economic theorizing, namely that the assumptions diverge in essential respects from the facts of real economic situations" (p. 586).

[3] Edelson gives no indication that he was aware of Buchanan's earlier work on this problem.

Otherwise, over- or undercongestion would result.

The Buchanan and Edelson qualifications are different but equally sufficient conditions for Knight's result. In this paper, a more general sufficient condition is presented that encompasses these as special cases. Nevertheless, the conclusion established by these authors, that optimal congestion under private ownership is rare, is not reversed.

In Sections I and II, a model of congestion-prone facilities is presented. It is more general in terms of the variety of institutional arrangements represented than that of the earlier contributors. One of its features is explicit inclusion of congestion effects in the facility demand function. This simplifies analysis by eliminating the separate "external cost" function frequently seen in models of externality problems (see, for example, A. Myrick Freeman and Robert H. Haveman).

Sections III, IV, and V compare the facility use levels under each ownership regime to the optimum. Aside from presenting the sufficient condition mentioned above, conditions leading to over- and undercongestion with private ownership are isolated.

Section VI compares the efficiency of the two regimes with each other and presents "qualitative," theoretical conditions associated with the superiority of each. Section VII summarizes results and mentions some policy implications.

I. The Congestion-Prone Facility

Let us begin with a congestion-prone facility and a set of N agents, the pool of would-be users. The service rendered by the facility is exactly the same for every simultaneous user. There is no variability in the intensity of individual use—one is either granted access to the facility or excluded.[4] Because of this, the number of *uses* of the facility is equal to the number of *users*. This number will be called x. In the road example, x is the number of drivers using the road; in the hunting preserve example, it is the number of hunters, etc. Due to congestion, the quality of the facility's service deteriorates as x increases.

To use partial equilibrium analysis, assume that all prices other than that of facility access are fixed at competitive levels. The most the ith agent will pay for access to the facility when there are x users is the reservation price $R^i(x)$. This is the compensating increment of income that leaves him indifferent between gaining access and being excluded when other prices are fixed as mentioned. Since the facility's service deteriorates in quality as x increases and congestion worsens, $R^i(x)$ is strictly decreasing in x for all i.

Where N is sufficiently large and the R^i functions sufficiently dense, x can be made continuous and aggregate demand for facility access can be approximated with a continuous function. Let g be a continuous, nonnegative variable indicating the level of congestion in the facility, and let $\Pi(x, g)$ be the twice differentiable function indicating the most that will be paid for the xth admission when the congestion level is g. The essential properties of the Π function are:[5]

$$\Pi_x(x,g)\leq 0, \Pi_g(x,g)<0 \qquad \text{for all } x\in[0,N] \text{ and } g\geq 0$$

Now let g be an explicit function of x, $g(x)$, where

$$g_x(x)>0 \text{ for all } x\in[0,N]$$

The single variable, inverse demand function for facility access can now be written as $\Pi(x, g(x))$. I define

$$x^c=\{x: \Pi(x,g(x))=0\}$$

to be the number of facility users when access is free, and impose the assumption that $x^c < N$. This means that when access is free, some

[4] The assumption that an agent's use of the facility is a 0–1 variable is not strictly necessary. Multiple, discrete uses or a continuous use variable could have been allowed without doing violence to essential results. The present assumption is used to ease the interpretation of analytical results, by emphasizing *users* rather than *uses*, and to preserve consistency with the Knight-Buchanan-Edelson framework.

[5] Subscripts are used throughout to designate partial derivatives.

agents who might otherwise desire it will refuse it due to excessive congestion. It follows from these assumptions that Π is a decreasing, one-to-one function that maps $[0, x^c]$ onto $[\Pi(0, g(0)), 0]$.

To round out the model I impose two further assumptions. First, for facilities where some notion of a fixed capacity is appropriate (for example, a museum), it is assumed to be sufficiently great to accommodate x^c simultaneous users. And second, it is assumed that all facility construction and maintenance costs are fixed, sunk, or zero; the marginal (noncongestion) cost of use is zero. The "sufficient capacity" assumption is relaxed somewhat later and some implications explored. The "zero-marginal cost" assumption is a convenient formalization of the observation that fixed costs are far more important than marginal costs for all congestion-prone facilities of the type mentioned previously.

II. Optimal Congestion

In what follows, welfare statements are made using a consumer's surplus approach. Because of the congestion externality, the area under the demand function

$$\int_0^x \Pi(y, g(y))\,dy$$

does not represent aggregate benefits when there are x users of the facility. The proper measure is computed by holding g at the actual level realized rather than letting it vary with the variable of integration:

$$(1) \qquad B(x) = \int_0^x \Pi(y, g(x))\,dy$$

This avoids attributing "phantom" benefits to inframarginal users.

The Π function in (1) is not, strictly speaking, an inverse demand function since g is parameterized. Instead, I call it a *benefit function*. It indicates reservation prices for facility access when the congestion level is $g(x)$. There is such a function for every g, and thus every x. The relationship between benefit functions and the inverse demand function is illustrated in Figure 1. The demand function is the locus of points on all benefit functions where x equals the level of use that defines them.

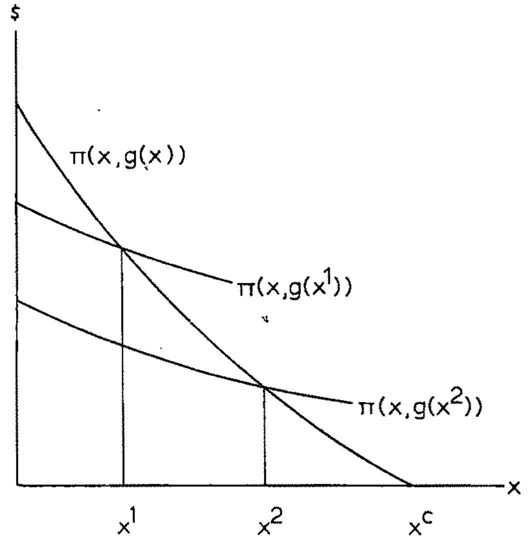


FIGURE 1

The optimal level of use for the facility, x^*, is one that maximizes $B(x)$ over x. Assume that $B(x)$ is strictly concave:

$$B_{xx}(x) < 0 \text{ for all } x \in [0, N]$$

This means that the necessary and sufficient first-order condition for a maximum is

$$(2) \quad \Pi(x^*, g(x^*)) + \int_0^{x^*} \Pi_g(x, g(x^*)) \cdot g_x(x^*) \cdot dx = 0$$

It also means that x^* is unique. To achieve x^* and the optimal congestion level, $g(x^*)$, an access price P^* must be charged. From (2),

$$(3) \quad P^* = -\int_0^{x^*} \Pi_g(x, g(x^*)) \cdot g_x(x^*) \cdot dx$$

It is the marginal social cost of an additional use, the sum of benefit reductions suffered by all users when another user is added.

III. Common Ownership

Suppose the facility is jointly owned by the N agents and regarded as common property. In the view of J. H. Dales, this regime of ownership is "virtually non-ownership" (p. 795). When everyone owns the facility, no one has the right to charge an access fee and exclude those unwilling to pay. Barring some kind of binding collective agreement to the contrary, access is free and the use level is x^c.

To evaluate the common ownership arrangement on grounds of economic efficiency, x^c is compared to x^*. From (2) it follows that $\Pi(x^*, g(x^*))>0$. Also, from previous assumption, $\Pi(x^c, g(x^c))=0$. Together these imply the familiar result:

PROPOSITION 1: $x^*<x^c$. *The facility is overused and congestion is excessive when it is regarded as common property.*

IV. Private Ownership

Suppose the facility is privately owned now, and the owner, a nondiscriminating monopolist, charges every user the same access price. This price, P^p, will be chosen to maximize revenues since all costs are fixed, sunk, or zero. (It is also assumed that costs incurred in levying the charge are nil.) The facility use level that corresponds to P^p is

$$x^p=\left\{x\colon \max_{(0,x^c)}\left[\Pi(x, g(x))\cdot x\right]\right\}$$

Assume that the revenue function is strictly concave:

$$d^2(\Pi(x, g(x))\cdot x)/dx^2<0 \text{ for all } x\in[0, N]$$

This means that the necessary and sufficient first-order condition for x^p is

$$(4)\quad \Pi(x^p, g(x^p))+\left[\Pi_x(x^p, g(x^p)) +\Pi_g(x^p, g(x^p))\cdot g_x(x^p)\right]\cdot x^p=0$$

and that x^p is unique. Thus:

$$p^p=-\left[\Pi_x(x^p, g(x^p)) +\Pi_g(x^p, g(x^p))\cdot g_x(x^p)\right]\cdot x^p$$

This price depends only on the magnitude of the marginal revenue loss suffered by the facility owner when an additional user is admitted and inframarginal users pay less. Unlike P^*, it does not depend explicitly on the marginal social cost of congestion and so does not accurately reflect the damage borne by inframarginal users when an additional user is admitted.

To evaluate this ownership arrangement, x^p must be compared to x^*, or equivalently, P^p to P^*. To do this, I explicitly define:

$$MSC(x)=-\int_0^x \Pi_g(y, g(x))\cdot g_x(x)\,dy$$

$$MRL(x)=(-\Pi_x(x, g(x)) -\Pi_g(x, g(x))\cdot g_x(x))\cdot x$$

to be marginal social cost and marginal revenue loss coincident with the admission of the xth facility user, $x\in[0, N]$. Both expressions are nonnegative. The private ownership counterpart to Proposition 1 is less conclusive.

PROPOSITION 2: *$P^*\lesseqgtr P^p$ and $x^*\gtreqless x^p$ if and only if $MRL(x^p)\gtreqless MSC(x^p)$.*

PROOF:

Equation (2) can be written as $\Pi(x^*, g(x^*))-MSC(x^*)=0$, which implies that

$$(5)\quad \Pi((x, g(x))-MSC(x)\gtreqless 0 \text{ iff } x\lesseqgtr x^*$$

Equation (4) can be rewritten as $\Pi(x^p, g(x^p))-MRL(x^p)=0$, from which it follows that

$$(6)\quad \Pi(x^p, g(x^p))-MSC(x^p)\gtreqless 0 \text{ iff } MRL(x^p)\gtreqless MSC(x^p)$$

Equations (5) and (6) imply together that

$$x^*\gtreqless x^p \text{ iff } MRL(x^p)\gtreqless MSC(x^p)$$

which establishes the result.

According to this proposition, whether private ownership leads to optimal congestion depends on whether the market signals provided by the marginal user (for those are all the facility owner cares about) are representative of inframarginal users' aversion to congestion. In general, they will not be and over- or undercongestion will occur.

To see the range of possibilities more clearly, let us differentiate cases according to the sign of $\Pi_{gx}(x, g)$.[6] To begin, suppose users with low reservation prices for facility access are systematically and consistently more sensitive to congestion than those with higher reservation prices:[7]

$$(7)\quad \Pi_{gx}(x, g)<0 \quad \text{for all } x\in[0, N],\ g\geqslant 0$$

This means that

$$(8)\quad -\Pi_g(x^p, g(x^p))\cdot g_x(x^p)> -\frac{1}{x^p}\int_0^{x^p}\Pi_g(y, g(x^p))\cdot g_x(x^p)\,dy$$

Thus $MRL(x^p)>MSC(x^p)$ unambiguously, and from Proposition 2, undercongestion is certain.

A second possibility is that users with high reservation prices are more sensitive to congestion than those with lower reservation prices:

$$(9)\quad \Pi_{gx}(x, g)>0 \text{ for all } x\in[0, N],\ g\geqslant 0$$

On intuitive grounds alone, this seems a more plausible hypothesis than (7) since it allows for congestion effects that are proportional to user benefits. In this case, however, the inequality in (8) is reversed and nothing conclusive about the relative magnitude of $MRL(x^p)$ and $MSC(x^p)$ can be inferred. The facility can, in principle, be over-, under-, or even optimally congested.

Pressing this case somewhat further, the price elasticity exhibited by the demand function at any x can be divided into two components (whose reciprocals are additive). The first component is the elasticity embodied in the operative benefit function:

$$(10)\quad \eta_1(x)=\frac{-\Pi(x, g(x))}{x\cdot\Pi_x(x, g(x))}$$

It varies inversely with the disparity of users' reservation prices for access when congestion is held constant. It does not reflect marginal congestion effects. The second is the elasticity produced by congestion effects alone:

$$(11)\quad \eta_2(x)\frac{-\Pi(x, g(x))}{x\cdot\Pi_g(x, g(x))\cdot g_x(x)}$$

It varies inversely with the sensitivity of users' reservation prices to facility congestion. If (9) holds, $MRL(x^p)>MSC(x^p)$ is more likely the greater is $-\Pi_x(x^p, g(x^p))$ or, equivalently, the smaller is $\eta_1(x^p)$. It follows from Proposition 2, then, that undercongestion is more likely the smaller is $\eta_1(x^p)$, or the less elastic are benefit functions in general. Conversely, as $\eta_1(x^p)$ increases, so does the likelihood that overcongestion occurs. This conclusion can be restated another way. Private ownership is more likely to under- (over-) congest the facility the more disparate (uniform) are user reservation prices for access at any constant congestion level.

The last case to be considered is where users are equally sensitive to congestion:

$$(12)\quad \Pi_{gx}(x, g)=0 \text{ for all } x\in[0, N],\ g\geqslant 0$$

Here, the inequality in (8) is changed to equality and $MRL(x^p)\geqslant MSC(x^p)$. Unless $-\Pi_x(x^p, g(x^p))=0$ so that $\eta_1(x^p)$ is infinite and benefit functions are horizontal, private ownership leads to under congestion. Instances where benefit functions are horizontal are taken up next.

V. The Knight-Buchanan-Edelson Propositions

Knight believed optimal congestion in congestion-prone facilities could be assured by merely establishing private ownership rights over them. Buchanan and Edelson

[6] This procedure does not exhaust the possibilities but is illustrative of their range. The delineation of cases along similar lines is suggested by A. Michael Spence.

[7] If the assumption that agent's use of the facility is a 0–1 variable were replaced with the assumption that it is a continuous, non-negative variable, then the interpretation of (7) would be that low-value uses are more sensitive to congestion than high-value uses where each use is identified with a specific user, but each user with many uses. The line of interpretation throughout the rest of Section IV would have to be altered accordingly. See fn. 4.

showed independently that this proposition is false for most institutional contexts. It is clear from Proposition 2 and the analysis above that Knight was mistaken. The present concern is to isolate theoretical conditions that do insure optimal congestion under private ownership.

From Proposition 2, a necessary and sufficient condition for optimal congestion is that $MRL(x^p)=MSC(x^p)$. Although there are others,[8] the most apparent case where this condition holds is the following special case of (12):

$$(13)\quad \Pi_x(x,g)=0 \text{ for all } x\in[0,N],\ g\geqslant 0$$

Equation (13) means that $\eta_1(x)$ is infinite for all $x\in[0,N]$ and that benefit functions are horizontal. All users share a common reservation price for access although this price may vary with the level of facility congestion. If the demand curve is downward sloping over $x\in[0,N]$, it is only because of pure congestion effects. Together with Proposition 2, (13) yields:

PROPOSITION 3: *If* (13), *then* $P^*=P^p$ *and* $x^*=x^p$.

Condition (13) is a generalization of the sufficiency conditions for optimal congestion supplied by Buchanan and Edelson. Buchanan's condition was that the facility have a sufficient number of competitors that the owner has no monopoly power. This condition satisfies (13) because it guarantees that users have a common reservation price: the lowest price asked by a competitor (all competitors, in equilibrium). The essential implication of his condition, taking it in the present context, is that benefit functions are horizontal. But this could be true for other reasons and Edelson's condition is an example. He required that road users in a modified Pigou-Knight example be homogeneous in their valuation of time. In this example, different time values is the only possible source of heterogeneity in reservation prices barring changes in the congestion level. Hence, the essential implication of Edelson's condition is also (13).

So (13) can hold for one of several reasons depending upon the institutional arrangements at hand. In some circumstances (for example, multiple lakes) it can hold because of an abundance of substitute facilities. In others (for example, an urban neighborhood park) it can hold because facility users are homogeneous in some essential way. Whether benefit functions are horizontal for one or the other of these reasons makes no difference. Both are comparatively rare, but either is sufficient.[9]

VI. Comparison of Ownership Arrangements

Apart from these special cases, neither ownership regime promises an optimally congested facility. Yet nothing has been said about the second best question as to which of them is better. Depending on demand conditions, either can be.

The possibilities are illustrated in Figure 2. The $B(x)$ function in the figure is strictly concave as assumed earlier, and x^* and x^c are identified. Although, following Proposition 2, the location of x^p vis-à-vis x^* is ambiguous, (4) implies that

$$(14)\qquad x^p<x^c$$

Private ownership always leads to a lower level of use than common ownership.

The first case to consider is where private ownership leads to optimal or overconges-

[8] For instance, subintervals of x on $[0,N]$ can exist where $\Pi_{gx}(x,g)\lesseqgtr 0$.

[9] The assumption at the beginning of Section IV that the facility owner charges every user the same price is crucial. Should the owner be a perfect price discriminator, he would levy a set of individual charges that would appropriate the entire users' surplus (the area under the benefit function) and would produce an optimal level of facility congestion. Indeed, this is the direction in which one must move in order to rescue the generality of Knight's proposition. There are sound reasons, however, for doubting that a facility owner will have the considerable information needed to be a perfect price discriminator. The reasons are mentioned in Section VII below in connection with the formulation of optimal social policy.

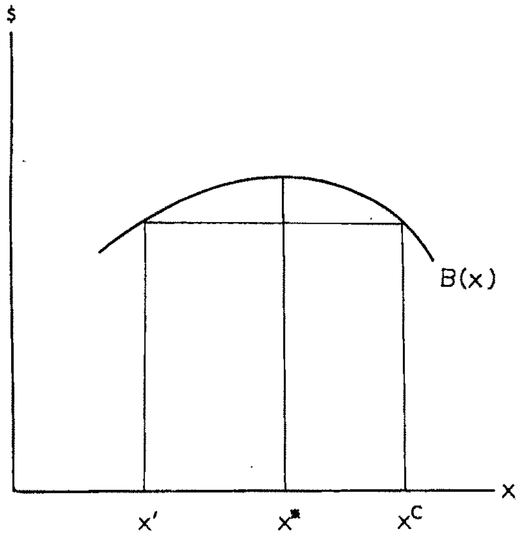


FIGURE 2

tion. A straightforward implication of either of these outcomes together with (14) is

PROPOSITION 4: *If* $x^p \geqslant x^*$, *then* $B(x^p) > B(x^c)$.

The extent of overcongestion is never as great with private as with common ownership. Nor is the associated welfare loss. Thus, knowledge that private ownership leads to overcongestion is all that is needed to establish its superiority over the alternative regime. Stated differently, Proposition 4 implies that common ownership can be superior only where private ownership would leave the facility undercongested.

Identifying the superior regime is more difficult when private ownership leads to undercongestion. The possibilities for this case are delineated as follows. Define x' in Figure 2 as $x' = \{x: x < x^* \text{ and } B(x) = B(x^c)\}$. From this we see $B(x^p) \gtreqless B(x^c)$ iff $x^p \gtreqless x'$. Private ownership is superior when the damage it causes from undercongestion is less than the damage caused by overcongestion under common ownership and conversely.

The undercongestion damages of private ownership and the overcongestion damages of common ownership arise for different reasons. The *comparative advantage of private ownership* is that congestion effects are internalized in the owner's revenue-maximizing calculus. The magnitude of this advantage is inversely related to the size of $\eta_2(x)$. As $\eta_2(x)$ decreases for all relevant x, users' reservation prices become more sensitive to changes in congestion and the damage (avoided by private ownership) from overcongestion increases. The *comparative advantage of common ownership* is that monopoly power, present when (13) does not hold and users have heterogeneous (constant congestion) reservation prices, cannot be exploited. The size of this advantage is inversely related to that of $\eta_1(x)$. As $\eta_1(x)$ decreases, for all relevant x, the disparity in user's (constant congestion) reservation prices increases and the incentive of private owners to undercongest is enhanced. Common ownership eludes this incentive and prevents the attending welfare loss.

From this, it can be concluded that private property is more likely superior the greater is $\eta_1(x)$ relative to $\eta_2(x)$ and conversely for common ownership. An example is helpful in illustrating this point.

Consider the case of a linear inverse demand function with linear benefit functions. Let

$$\Pi(x, g(x)) = \alpha - \beta x - \gamma g(x), \quad \text{for all } x \in [0, N]$$

where $-\Pi_x(\cdot) = \beta \geqslant 0$, $-\Pi_g(\cdot) = \gamma > 0$, and let

$$g(x) = x \text{ for all } x \in [0, N]$$

This is an example of (12) where all users are equally sensitive to congestion. Solving for x^*, x^c and x^p yields $x^* = \alpha/(\beta + 2\gamma)$; $x^c = \alpha/(\beta + \gamma)$; $x^p = \alpha/2(\beta + \gamma)$. Comparing these, we find $x^p \leqslant x^* < x^c$. This is shown in Figure 3; common ownership leads to overcongestion and private ownership to undercongestion (if $\beta > 0$) or optimal congestion (if $\beta = 0$, in which case (13) holds).

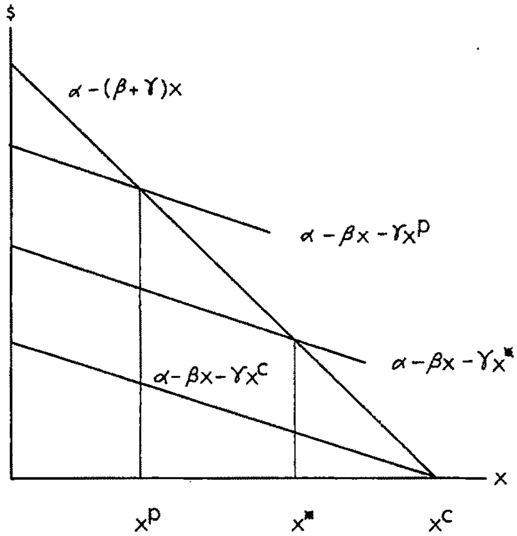


FIGURE 3

To compare aggregate benefits produced under each regime, we solve:

$$B(x^p)=\frac{\alpha^2}{(\beta+\gamma)}\left[1/2-\frac{(\gamma+\beta/2)}{4(\beta+\gamma)}\right]$$

$$B(x^c)=\frac{\alpha^2}{(\beta+\gamma)}\left[1-\frac{(\gamma+\beta/2)}{(\beta+\gamma)}\right]$$

From these it follows that

$$(15)\qquad B(x^p)\lesseqgtr B(x^c)\text{ iff }\gamma\lesseqgtr\beta/2$$

Recalling (10) and (11), $\eta_1(x)$ and $\eta_2(x)$ for this example are

$$\eta_1(x)=\frac{\alpha-\beta x-\gamma x}{\beta x}$$

$$\eta_2(x)=\frac{\alpha-\beta x-\gamma x}{\gamma x}$$

These with (15) yield

$$B(x^p)\lesseqgtr B(x^c)\text{ iff }\eta_1(x)$$

$$\lesseqgtr\eta_2(x)/2\text{ for any }x\in[0,x^c]$$

Private ownership is superior as long as $\eta_1(x)$ exceeds half of $\eta_2(x)$ for any (and thus all) relevant facility use level. Common ownership is superior if $\eta_2(x)$ is more than twice $\eta_1(x)$.

While this discussion of the relationship between the relative size of $\eta_1(x)$ and $\eta_2(x)$ and the identity of the superior ownership regime has been confined to the case where $x^p<x^*$, the conclusion above holds generally. From analysis in IV, $x^p\geqslant x^*$ was seen to be more likely the greater is $\eta_1(x^p)$; from Proposition 4 it follows that private ownership superiority is also more likely. Recalling the interpretations of $\eta_1(x),\eta_2(x)$, this permits the stating of the following general proposition.

PROPOSITION 5: *The private property regime is more likely superior the more uniform are users' (constant congestion) reservation prices, and the more sensitive these prices are to congestion changes. The converse holds for the common property regime.*

This proposition suggests qualitative criteria for judging the relative superiority of the two ownership arrangements in specific institutional contexts. For facilities having substantial congestion effects and either largely homogeneous users or an abundance of substitutes, the private ownership arrangement is preferred. For those where congestion effects are minimal, but users heterogeneous and substitutes absent, common ownership is better. For cases in between, qualitative criteria alone offer little guidance in determining which regime is more efficient. For these it is necessary to specify functional forms for $\Pi(\cdot)$ and compute $B(x^p)$ and $B(x^c)$ in order to make judgements. The appeal of the qualitative criteria in Proposition 5 is enhanced by the observation, elaborated upon in the final section, that computation of $B(\cdot)$ is usually infeasible for informational reasons.

A final consideration in comparing ownership arrangements is the effect of a strict capacity constraint on the facility. Suppose, contrary to previous assumption, that the facility cannot physically accommodate x^c simultaneous users but only $\bar{x}$, where $\bar{x}<x^c$ and $\Pi(\bar{x},g(\bar{x}))>0$. (There is bound to be an $\bar{x}$ for almost any congestion-prone facility.

What makes this case different is that $\bar{x}<x_c$ so that the capacity constraint is active under common ownership.)

The main difference here is that $B(\bar{x})$ is an imperfect measure of aggregate benefits in the common ownership regime. Since $\bar{x}<x^c$, excess demand exists under free access and some form of nonprice rationing must arise to determine which agents are admitted. If the rationing is efficient[10]—that is, if access is provided for those $\bar{x}$ users with the highest reservation prices—then aggregate benefits are $B(\bar{x})$. If rationing is inefficient but imposes no (pecuniary or other) access charge on those admitted, benefits can be as small as

$$(16) \qquad \int_{x^c-\bar{x}}^{x^c} \Pi(x, g(\bar{x}))\,dx$$

If the rationing scheme employed involves a nonpecuniary access charge (as for instance with rationing by queuing where the charge depends on the imputed values of time spent waiting), then a true measure of benefits is more difficult to provide. If the effective "price" of access varies among users (and is not a mere transfer among agents), it is possible that benefits are actually less than (16).

VII. Conclusions and Implications

There are three principal conclusions of the paper.

(i) Common ownership of congestion-prone facilities always leads to overcongestion (Proposition 1) while private ownership leads variously to over-, under-, or optimal congestion depending on whether and how market signals provided by the marginal user misrepresent inframarginal users' aversion to congestion (Proposition 2). If users with low reservation prices for access are more sensitive to congestion than those with higher ones, undercongestion occurs with private ownership. If users with high reservation prices are more sensitive, the result is indeterminate a priori.

(ii) Private ownership leads to optimal congestion only where market signals provided by the marginal user are representative of inframarginal users' aversion to congestion. A necessary condition for this is that the marginal revenue loss to the facility owner of the x^pth user be equal to the marginal social cost produced by that user. A stronger and sufficient condition is that users' benefit functions be horizontal or infinitely elastic (Proposition 3). Depending on the facility in question, this can occur either because there is an abundant supply of substitute facilities, or because users are homogeneous in some relevant way.

(iii) Comparing the ownership arrangements to each other, private ownership is unambiguously superior when it leads to overcongestion (Proposion 4). More generally, the likelihood that private ownership is superior increases as disparity in users' (constant congestion) reservation prices decreases and as these prices become more sensitive to changes in the level of congestion (Proposition 5).

This last conclusion has policy implications. There is a rich literature on the question of optimal social policy for externality problems like congestion-prone facilities. If a facility is common property, the standard welfare-economic prescription would be for government to charge an access fee. If it is private property, the government should intervene as a price regulator. In either case, the optimal price is upheld and the facility is optimally congested.

However sound in principle, these recommendations are threatened by a formidable problem once implementation begins. The information needed to compute P^* is encoded in the benefit functions and not in the ordinary demand function for the facility. The demand function (or its relevant segment) is observable and can be reconstructed from market data. But benefit functions are not observable except at the one point each shares with the demand function. It is not possible to reconstruct them (or even local segments of them to make marginal evaluations) using market data. To see this, note that for every demand function revealed by observation or experiment there is an infinite number of sets of benefit functions that are compatible. To choose one over the others

[10] This is the terminology used, for instance, by Roger Sherman and Michael Visscher.

(and disregarding the complication of higher order, non-linear effects) we must decompose the slope of the demand function at x into $\Pi_x(x, g(x))$ and $\Pi_g(x, g(x))\cdot g_x(x)$, or equivalently, total price elasticity into $\eta_1(x)$ and $\eta_2(x)$. Market information alone provides no help in such a decomposition. For this reason, *optimal* social policy is infeasible.

It was in a context similar to this that William Baumol endorsed a second best approach: "[G]iven the limited information at our disposal, it is perfectly reasonable to act on the basis of a set of minimum standards of acceptibility" (p. 318). An example of this kind of approach is offered by Dales in his "transferable pollution rights" scheme for regulating water quality in a common-property water resource. He envisaged government choosing an acceptible level of water quality, creating a corresponding number of pollution rights (each being a permit for a specific "amount" of water pollution), and auctioning off these rights to private water-users. While the level of water quality chosen would probably not be optimal, at least the allocation of scarce polluting capacity among users would be efficient.

With the problem at hand, the Baumol-Dales approach would be to achieve efficient rationing of access subject to an "acceptable" level of facility congestion. This prescription is both realistic and promising. But once we have retreated this far from the optimum it is no idle question to ask whether one of the ownership arrangements discussed in this paper will not provide an equally acceptable congestion level.[11] This is certainly the case for a facility conforming to one of the qualitative extremes described above in Section VI.

In this connection, Proposition 5 suggests that welfare gains are sometimes achievable by simply changing a facility's ownership arrangement. By exercising its eminent domain powers and making restitution out of general revenues, government can convert privately owned facilities to common property when the latter is preferable. In other circumstances, common property facilities can be converted to private property by auctioning them off to the highest bidder and distributing the proceeds by reducing general tax bills. If "franchise bidding" is sufficiently competitive, the entire surplus extracted by the winner will be transferred to "common" owners (see Harold Demsetz). In either case the congestion level is set by market forces, as government is given no role in price or congestion-standard setting.

A thorough analysis of the merits of such changes in ownership arrangements would require comparison with other second-best candidates. This comparison would necessarily involve policy administration costs and vulnerabilities to political malfeasance. For these reasons and others, it is not undertaken here.

REFERENCES

W. J. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–22.

J. M. Buchanan, "Private Ownership and Common Usage: The Road Case Reexamined," *Southern Econ. J.*, Jan. 1956, *22*, 305–16.

J. H. Dales, "Land, Water, and Ownership," *Can. J. Econ.*, Nov. 1968, *1*, 791–804.

H. Demetz, "Why Regulate Utilities," *J. Law and Econ.*, Apr. 1968, *11*, 55–66.

N. M. Edelson, "Congestion Tolls Under Monopoly," *Amer. Econ. Rev.*, Dec. 1971, *61*, 873–82.

A. M. Freeman and R. H. Haveman, "Congestion, Quality Deterioration, and Heterogenous Tastes," *J. Public Econ.*, Oct. 1977, *8*, 225–32.

H. S. Gordon, "The Economic Theory of a Common-Property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124–42.

F. H. Knight, "Some Fallacies in the Interpretation of Social Cost," *Quarterly J. Econ.*, Aug. 1924, *38*, 582–606.

A. C. Pigou, *The Economics of Welfare*, London 1920.

R. Sherman and M. Visscher, "Second Best Pricing with Stochastic Demand," *Amer. Econ. Rev.*, Mar. 1978, *68*, 41–53.

A. M. Spence, "Monopoly, Quality and Regulation," *Bell. J. Econ.*, Autumn 1975, *3*, 417–29.

[11] These, recall, lead to efficient rationing in all cases but the one mentioned above in Section VI where $\bar{x}<x^c$.

Capacity, Output, and Sequential Entry

By DANIEL F. SPULBER*

Two crucial assumptions frequently made in industrial organization are that an established firm deters entry either by a constant high output (the Sylos Postulate) or by high excess capacity (the Excess Capacity Hypothesis). These assumptions may be an accurate description of the observed conduct of firms in some industries. However, when the rules of the post-entry game are clearly specified, the optimal output and investment strategies of an established firm may depart considerably from these *behavioral* assumptions. Because of this possible inconsistency with rational behavior, any conclusions about the formation of industry based upon either assumption are highly suspect. What is more, these assumptions avoid the main issue of whether entry deterrence is worthwhile at all.

This paper presents a dynamic model of entry in which established firms pursue a Cournot-Nash (alternatively Stackelberg) strategy toward a potential entrant. The entrant behaves in Cournot-Nash fashion and chooses output on the basis of *expected post-entry profits at the equilibrium of the post-entry game*. Within this framework, a constant output entry-deterring strategy would involve maintenance of an entry-deterring output level before and after entry is threatened. An excess capacity entry-deterring strategy would involve holding excess capacity at an entry-deterring level and increasing output to that level after entry is threatened. Special conditions are presented under which the Sylos Postulate or the Excess Capacity Hypothesis will accurately describe optimal entry-deterring strategies. In addition, special conditions are examined under which the established firm maintains a constant output or holds pre-entry excess capacity when large-scale entry does in fact take place. The analysis shows that in general, established firm reactions to entry are quite different from these special cases.

The Sylos Postulate (see Joseph Bain; Paolo Sylos-Labini; Franco Modigliani, 1958) asserts not only that potential entrants expect established firms to maintain their output constant as entry occurs, but that established firms keep output constant at a level that deters entry *whether or not it is profitable to do so*. Hailed as a "welcome major breakthrough on the oligopoly front" (Modigliani, 1958) the Sylos Postulate underlies many papers in the large theoretical and empirical literature on limit pricing.[1] Yet the Sylos Postulate ignores both the strategic interaction between firms and the dynamic aspects of entry.[2] Unless the established firm's monopoly output exceeds the entry-deterring level, the established firm with a general cost function able to choose an entry-deterring output level will instead *always desire a lower output level before entry*. For the special case where capacity is an upper bound on output and the established firm is a Stackelberg leader in the post-entry game,

*Assistant professor of economics, Brown University. The support of the National Science Foundation, under grant No. SOC79-07201 is gratefully acknowledged. I would like to thank Robert Becker for helpful discussions. The paper has benefited from very useful suggestions from George Borts, Avinash Dixit, and an anonymous referee. Any errors are my own responsibility. I also thank Marion Wathey for expert typing.

[1] For a discussion and additional references see Dale Osborne (1964), Jagdish Bhagwati, F. M. Scherer (pp. 219–34), and Douglas Needham (ch. 7).

[2] The point made here is quite different from that made by Franklin Fisher and echoed by Modigliani (1959) who find that Cournot and Sylos behavior simply imply different *outcomes* in a market model. The problem lies in finding a specification of the rules of the post-entry game in which a constant output is a rational strategy for the established firm. Employing alternative *ad hoc* assumptions concerning the behavior of entrants, Peter Pashigian and John Wenders (1971b) find that an established firm may vary its output over time as a profit-maximizing response to entry. In a *static* model, Avinash Dixit (1980) examines a Nash equilibrium in the post-entry game and an equilibrium where the *entrant* is a Stackelberg leader and finds that the excess capacity strategy will not be employed. This paper goes beyond Dixit's by introducing a dynamic analysis, by employing a more general cost function and by allowing the *established firm* to be a Stackelberg leader.

this paper presents a necessary and sufficient condition for the firm to maintain a constant entry-deterring output level. These results imply that models which depend upon the Sylos Postulate, such as static limit pricing, may be quite misleading. By allowing some entry to occur, Darius Gaskins and Morton Kamien and Nancy Schwartz (1971) have studied dynamic limit pricing, yet their *ad hoc* assumption of entry as a function of current market price is inconsistent with expected profit-maximizing behavior by entrants. This paper demonstrates that *when entry does take place, a necessary and sufficient condition for the established firm to maintain a constant output is a high cost of capacity relative to the net discounted marginal returns.*

The Excess Capacity Hypothesis, which allows the established firm to freely vary its output, was proposed as an alternative to the Sylos Postulate by Pashigian and Wenders (1971a) and extended by A. Michael Spence (1977) and Dixit (1979).[3] Spence assumes that potential entrants base their entry decision on the capacity of the established firm and he then constrains the established firm to choose capacity at or above the entry-deterring level *whether or not it is profitable to do so.* Thus the Excess Capacity Hypothesis also ignores the strategic interaction between firms and the dynamic aspects of entry. In particular, it fails to recognize that the established firm's choice of costly capacity involves a tradeoff between pre- and post-entry requirements. Spence's conclusion (1977) that the pre-entry price may exceed the limit price and that the pre-entry quantity supplied by the established firm may be less than the limit quantity is certainly not surprising in view of the requirement that capacity be at an entry-deterring level.

This paper shows the Excess Capacity Hypothesis to be inconsistent with post-entry Cournot-Nash behavior whether or not entry is permitted by the established firm. *Holding excess capacity to deter entry is shown to occur only when the established firm is a Stackelberg leader, the Stackelberg output exceeds the short-run monopoly output and the cost of capacity is low relative to net discounted marginal returns at the entry-deterring output.* This paper shows that the necessary and sufficient condition for entry deterrence with pre-entry excess capacity is that the discounted sum of marginal post-entry profits and the marginal value of the entrant's reaction evaluated at the entry-deterring output must equal or exceed the cost of capacity.

Section I presents the capacity investment framework. The Cournot-Nash case is examined with capacity as an upper bound on output in Section II and with a general cost function in Section III. The Stackelberg case is examined for the special cost function in Section IV and for the general cost function in Section V.

I. The Capacity Investment Framework

Capacity investment by a firm established in a particular industry depends primarily on the firm's strategic response over time to actual or potential entry. An established firm with a given plant size experiencing demand fluctuations due to entry, may have excess capacity either before or after entry. The dynamic aspects of capacity and imperfect competition, are discussed extensively by Roy Harrod, John Hicks, and Frank Hahn, who emphasize the Marshallian distinction between the short and long periods. Hicks and Hahn distinguish between a *closed period* when a monopolist completes construction of his plant and begins producing output and an *open period* when competitors have had time to construct similar plants and also begin supplying output. Arguing that potential entrants "take account not of the actual profits of existing producers but of the profits they themselves could earn if they entered" (p. 239), Hahn finds that given a downward sloping demand curve the established firm may hold excess capacity after entry takes place. Similar results are obtained by Kamien and Schwartz (1972) who examine dynamic capacity investment when the threat of entry is a random function of price. Kamien and Schwartz find that the

[3] When fixed costs are present, Dixit (1979) finds that the Excess Capacity Hypothesis allows the established firm to block entrants who could not be effectively deterred with a conventional limit pricing strategy.

relative effects of pre-entry versus post-entry profit on the plant size chosen by a monopolist depend upon the rate of interest and on the riskiness of rival entry, but as in Harrod, Hicks, and Hahn, *they do not specify the strategy of the potential entrant or the post-entry reponse of the established firm.*

Dixit (1980) has examined the strategic role of investment in entry deterrence by specifying alternative rules for the post-entry game within a *static* framework. This paper goes beyond the analysis of Dixit by introducing an explicitly dynamic framework which takes into account entry lags.

The analysis presented here considers a two-period model of entry in which a monopolist chooses output and capacity in the first period with the knowledge that another firm will enter in the second period. This framework is chosen so as to emphasize the capacity investment decisions of an established firm over time when entry occurs sequentially. The model may be extended to the case where a group of firms acting in collusion faces the possibility of entry or to the case of entry of firms over several time periods. The analysis goes further than those of Hicks and Hahn by explicitly considering the strategies of the entrant and the established firm in the second period and examines their effect on the established firm's output and capacity choices in the first period. As in Hicks and Hahn the plant size chosen by the firm upon entering is permanent, reflecting the irreversibility of investment as well as imperfections in the rental market for capital. Capacity is purchased in a competitive market by a new firm and is an upper bound on the firm's output during any period. The cost function will be generalized in Section III.

Market demand is assumed to be the same in each period. In addition, it is also assumed that the inverse demand function $p(\cdot)$ is differentiable, decreasing and concave.

Let firm 1 be the established firm and firm 2 the potential entrant. Let x_1^1, x_2^1 denote the established firm's output in the first and second period, and let x^2 denote the entrant's output in the second period. Each firm purchases capacity k^1, k^2 at market price q. The variable costs of producing output in each period are given by $c^1(\cdot), c^2(\cdot)$. The variable cost functions $c^1(\cdot), c^2(\cdot)$ are assumed to be differentiable, increasing and convex. Let $c^1(0)=c^2(0)=0$.

It will be important to compare the entry problem to a monopoly facing no threat of entry. The monopolist will have the same capacity requirements in each period. Let $\pi(x)=p(x)x-c^1(x)$ denote the monopolist's short-run profit function in each period. Given an interest rate r, the monopolist chooses capacity k_m such that the discounted sum of marginal returns equals the cost of capacity q.

$$(1) \qquad \pi'(k_m)\left(1+\frac{1}{1+r}\right)=q$$

Note that since profits are concave and the cost of capacity is positive, capacity k_m is strictly less than the short-run profit-maximizing output M^1, which solves $\pi'(M^1)=0$. Given entry in the second period, this result will no longer hold. Let us now examine how the established firm's capacity and output choices are affected if the post-entry market has a Cournot-Nash or Stackelberg equilibrium.

II. The Cournot-Nash Case

The occurrence of large-scale entry in the second period is now introduced. Since the entrant produces for only a single period its output and capacity choice will be identical. The entrant's problem is to choose its output x^2 to maximize its profits net of capacity costs, given the second period output of the established firm x_2^1. The established firm begins the second period with a fixed plant capacity chosen when it entered the market in the first period. Thus, the established firm must choose its output x_2^1 subject to its capacity constraint $x_2^1 \leq k^1$ and taking the entrant's output x^2 as given. Therefore, given the established firm's capacity k^1, it can be shown that there exists a *capacity constrained Cournot-Nash equilibrium*[4] (x_2^{1*}, x^{2*}) where

[4] When capacity is held constant, static equilibrium in an oligopolist market (first studied by Augustin Cournot) may be viewed as the equilibrium solution to an

the outputs of the established firm and the entrant solve

$$(2)\qquad \max_{x_2^1}\left[p\left(x_2^1+x^{2*}\right)x_2^1-c^1\left(x_2^1\right)\right]$$

subject to $\qquad x_2^1 \leqslant k^1$

and

$$(3)\qquad \max_{x^2}\left[p\left(x_2^{1*}+x^2\right)x^2-c^2\left(x^2\right)-qx^2\right]$$

Consider the post-entry game where the capacity constraint on the established firm ($x_2^1 \leqslant k^1$) is *not* present. Let $x_2^1=\gamma^1(x^2)$ and $x^2=\gamma^2(x_2^1)$ represent the Cournot reaction functions for the established firm and entrant, respectively, in the *unconstrained* post-entry game. Let $M^1=\gamma^1(0)$ and $M^2=\gamma^2(0)$ be the *unconstrained monopoly outputs* and let (Q^1, Q^2) be the *entry blocking intercepts*, where $\gamma^1(Q^2)=0$ and $\gamma^2(Q^1)=0$. Note that the analysis is general enough to allow the entrant to have a discontinuous reaction curve due to high fixed costs or to a large minimum efficient scale relative to market demand.[5] Thus, in Figures 1 through 4, the *unconstrained* reaction curves are given by Q^2M^1 for the established firm and M^2Q^1 for the entrant.

I now turn to the *capacity-constrained* post-entry game. Given capacity k^1 the established firm's reaction is truncated at k^1. This implies that, unless the entry blocking output Q^1 is less than the short-run monopoly output of the established firm M^1, there is no possibility of the established firm deterring entry in the Cournot-Nash case. Even if $Q^1<M^1$, the established firm may *still* not find it worthwhile to deter entry when capacity is costly and may choose $k^1<Q^1$ (see Figure 1). Entry will be deterred if the monopolist's capacity level without threat of entry k_m equals or exceeds Q^1. This capacity investment condition may be interpreted as Bain's case of "blockaded entry" (see Bain, pp. 21–22, see also Dixit, 1979).

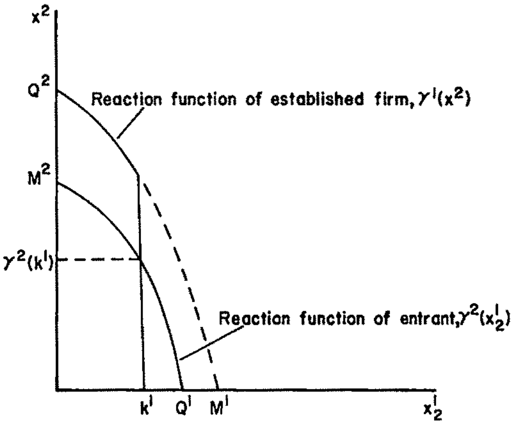

x^2
Q^2
Reaction function of established firm, $\gamma^1(x^2)$
M^2
$\gamma^2(k^1)$
Reaction function of entrant, $\gamma^2(x_2^1)$
k^1
Q^1
M^1
x_2^1

FIGURE 1. $Q^1<M^1$

PROPOSITION 1: *If the established firm follows a Cournot-Nash strategy, then entry will be blocked if and only if*

$$(4)\qquad \pi'(Q^1)\left(1+\frac{1}{1+r}\right)\geqslant q$$

The entry-deterring capacity level is chosen when interest rates are low or when the cost of capacity is low relative to discounted marginal returns at the entry-blocking output Q^1.[6] When the established firm follows a Cournot-Nash strategy in the post-entry game, the Sylos Postulate is only satisfied in this limited sense. Clearly, the established firm will produce at full capacity in each period.

n-person noncooperative game (see John Nash). The effect of entry on price and output when firms are following Cournot-Nash strategies was studied by Charles Frank, Roy Ruffin, and Koji Okuguchi who examined convergence to the competitive output as the number of firms increases. William Novshek demonstrates the existence and approximate competitiveness of a Cournot-Nash equilibrium with free entry in the presence of fixed costs. This is extended to a general equilibrium setting in Novshek and Sonnenschein (1978). See also Maurice McManus.

[5] The analysis includes the entrant's cost functions in Spence (1977) and Dixit (1979) as special cases. Dixit extensively examines the conditions where entry deterrence will be undertaken by the established firm when fixed costs are present (1979) and in a strategic investment framework (1980).

[6] This result is consistent with Spence who notes that "when demand is highly inelastic in the range of prices near marginal costs" (1977, p. 536, case 1), then unconstrained profit-maximizing decisions set capacity at the entry-deterring level.

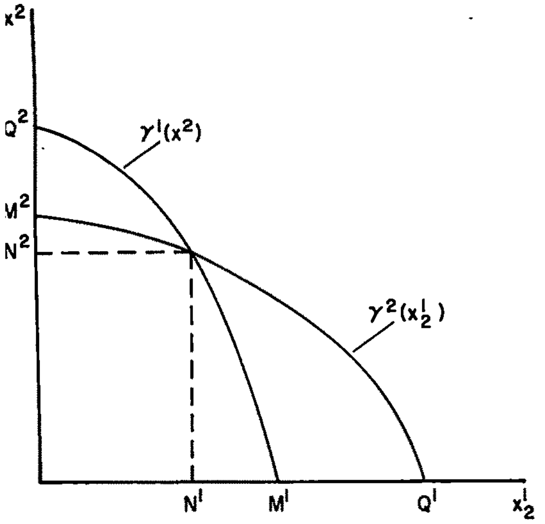


FIGURE 2. $Q^1>M^1$

Suppose now that the entry-blocking output Q^1 is strictly greater than the short-run monopoly output of the established firm M^1. Then, without the capacity constraint on the established firm, the post-entry game has a standard Cournot-Nash equilibrium (N^1, N^2) at the intersection of the reaction functions $(\gamma^1(x^2), \gamma^2(x_2^1))$ (see Figure 2). If capacity k^1 is chosen to be greater than or equal to N^1, then (N^1, N^2) will be the post-entry equilibrium. When capacity is strictly less than N^1, there is a capacity-constrained Cournot-Nash equilibrium at $(k^1, \gamma^2(k^1))$ (see Figures 3a and 3b).

Given that the established firm behaves competitively as outlined above, the Cournot-Nash equilibrium implicitly defines the value of capacity in the second period $V^N(k)$ given by

$$(5)\quad V^N(k^1) = \begin{cases} p(N^1+N^2)N^1-c^1(N^1)\text{if } k^1\geqslant N^1 \\ p(k^1+\hat{x}^2)k^1-c^1(k^1)\text{if } k^1\leqslant N^1 \end{cases}$$

where $\hat{x}^2=\gamma^2(k^1)$ is taken as given by the established firm. Note that $V^N(k^1)$ is differentiable in k^1 and is strictly concave for $k^1\leqslant N^1$. Also, for $k^1\geqslant N^1$, $V^{N'}(k^1)=0$.

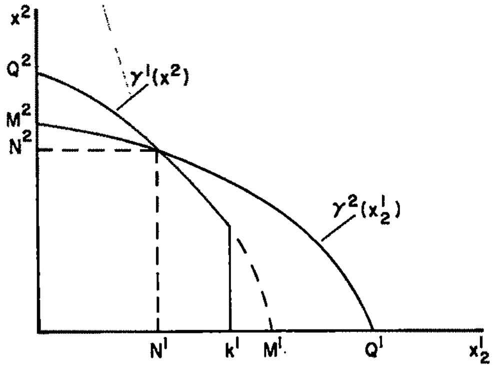


FIGURE 3a. $k^1\geqslant N^1$

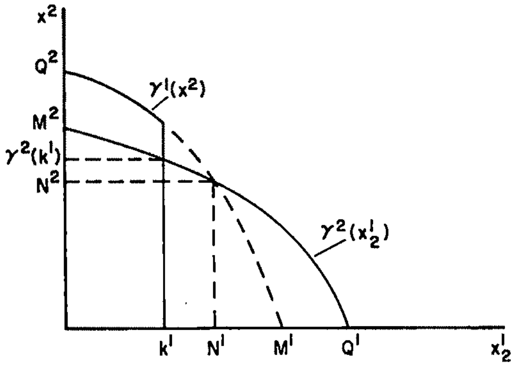


FIGURE 3b. $k^1<N^1$

Before entry, the established firm wishes to choose output x_1^1 and capacity k^1 to maximize the present discounted value of profits. Given the implicit value of capacity defined in (5), we may employ the backward induction approach of dynamic programming. Drop the firm superscripts and time subscripts. The established firm's problem is then

$$(6)\quad \max_{x,k}\left[\pi(x)-qk+\frac{1}{1+r}V^N(k)\right]$$

subject to $x\leqslant k$

The Kuhn-Tucker first-order necessary conditions for the problem are

$$(7)\quad \pi'(x)=\eta$$

$$(8) \qquad \eta+\frac{1}{1+r}V^{N'}(k)=q$$

$$(9) \qquad \eta(k-x)=0, \eta \geqslant 0$$

Analysis of conditions (7)–(9) yields the following result:

PROPOSITION 2: *If the established firm follows a Cournot-Nash strategy, then*

(a) *The firm operates at full capacity before entry takes place. Capacity is strictly less than* M^1.

(b) *Capacity is greater than, equal to or less than* N^1 *if and only if* $\pi'(N^1)$ *is greater than, equal to or less than* q.

PROOF:

(a) Suppose $\eta=0$. This implies that the constraint $x \leqslant k$ is nonbinding and from (7), $x=M^1$, so $M^1 \leqslant k$. If $k \leqslant N^1$ then since $N^1 < M^1$, there is a contradiction. If $k>N^1$, then $V^{N'}(k)=0$ which contradicts (8). So $\eta>0$, which implies that $x=k$. This also implies that $x<M^1$.

(b) If $k \geqslant N^1$ then $V^{N'}(k)=0$, so $\pi'(k)=q$. Thus if $k>N^1$, then $\pi'(N^1)>q$. If $k=N^1$, $\pi'(N^1)=q$. If $k<N^1$, then $\pi'(k)<q$, so that $\pi'(N^1)<q$. Since this exhausts the possible cases, the converse is also true.

The proposition shows that *holding excess capacity is inconsistent with a post-entry Cournot-Nash equilibrium.*[7] Further, *if the short-run net marginal return from producing at the Cournot-Nash output before entry does not exceed the cost of capacity, then output will be kept constant in the face of entry.* In other words, when capacity is relatively expensive, the established firm will produce at full capacity in each period at or below the Cournot-Nash output. In this case, capacity satisfies the optimality condition:

$$(10) \qquad \pi'(k)+\frac{1}{1+r}V^{N'}(k)=q$$

If the net marginal returns to producing at the Cournot-Nash output in the first period exceed the cost of capacity, the firm will initially produce at a higher level than the Cournot-Nash equilibrium. Therefore, *when capacity is relatively inexpensive, the established firm will contract its output in the face of entry, thus contradicting the Sylos Postulate.* Then the firm will operate at full capacity in the first period and carry excess capacity in the second. In this case capacity satisfies $\pi'(k)=q$.

III. General Cost Function

The pre-entry choice of irreversible capacity may be examined using the general cost function $c^1(x,k), c^2(x,k)$. This allows the established firm to equate marginal revenue to marginal cost after entry. It is assumed that $c^1(x,k)$ has the standard properties,[8] $c_x^1>0$, $c_{xx}^1>0$, $c_k^1<0$, $c_{kk}^1<0$, and $c_{kx}^1<0$. As before, capital k^1 is purchased at price $q>0$ when the firm is established.

The first-order condition for the established firm's second period output choice in the Cournot-Nash case is then

$$(11) \qquad p'(x_2^1+x^{2*})x_2^1+p(x_2^1+x^{2*}) = c_x^1(x_2^1,k^1)$$

This equation can be solved for the reaction function of the established firm $\gamma^1(x^2, k^1)$ which is downward sloping in x^2 and increasing in k^1. Since the entrant's reaction function is downward sloping in x_2^1, *an increase in* k^1 *increases the market share of the established firm in the second period.* The returns to increasing capital are, of course, limited by the Cournot-Nash strategy of accommodating entry.

Applying the envelope theorem, the marginal value of initial capital after entry is given by

$$(12) \qquad V^{N'}(k^1)=-c_k^1(x_2^1,k^1)$$

[7] A similar result is obtained by Dixit (1980, p. 100) in a static framework. The special case where excess capacity is held *after* entry is noted by Hahn, and Kamien and Schwartz (1972).

[8] Spence (1977) and Dixit (1980) employ similar cost functions. For a detailed derivation of a cost function of this type using a general production function, irreversible investment and adjustment costs, see my paper with Robert Becker.

Using (12), consider the first-order necessary conditions for the established firm's choice of output and capital before entry,

$$(13)\qquad p'(x_1^1)x_1^1+p(x_1^1)-c_x^1(x_1^1,k^1)=0$$

$$(14)\qquad -q-c_k(x^1,k^1)-\frac{1}{1+r}c_k^1(x_2^1,k^1)=0$$

The condition required for entry deterrence to be optimal is now examined. Substituting for $x_1^1=x_2^1=Q^1$ in (14) defines the optimal capacity level at the entry deterring output level $k(Q^1)$. The following condition insures that the monopoly output deters entry.

PROPOSITION 3: *If the established firm follows a Cournot-Nash strategy, then given the cost function $c^1(x,k)$ entry will be blocked if and only if*

$$(15)\qquad p'(Q^1)Q^1+p(Q^1)\geqslant c_x^1(Q^1,k(Q^1))$$

Consider the case where (15) is not satisfied and entry takes place. Differentiating (13) and (14) totally and applying Cramer's rule, I obtain some interesting comparative statics results. Let Δ be the determinant of the coefficient matrix, where Δ can be shown to be strictly positive thus satisfying the second-order sufficient conditions. The main results are

$$(16)\qquad \frac{dx_1^1}{dq}=\frac{1}{\Delta}c_{xk}^1(x_1^1,k^1)<0$$

$$(17)\qquad \frac{dk^1}{dq}=\frac{1}{\Delta}\big[p''(x_1^1)x_1^1+2p'(x_1^1)$$
$$-c_{xx}^1(x_1^1,k^1)\big]<0$$

$$(18)\qquad \frac{dx_1^1}{dr}=-\frac{1}{\Delta}\frac{1}{(1+r)^2}c_k(x_2^1,k^1)$$
$$\times c_{xk}^1(x_1^1,k^1)<0$$

$$(19)\qquad \frac{dk^1}{dr}=-\frac{1}{\Delta}\frac{1}{(1+r)^2}c_k^1(x_2^1,k^1)$$
$$\times\big[p''(x_1^1)x_1^1+2p'(x_1^1)-c_{xx}^1(x_1^1,k^1)\big]<0$$

$$(20)\qquad \frac{dx_1^1}{dx^2}=\frac{1}{\Delta}\frac{1}{1+r}\gamma_x^1(x^2,k^1)c_{kx}^1(x_2^1,k^1)$$
$$\times c_{xk}^1(x_1^1,k^1)<0$$

$$(21)\qquad \frac{dk^1}{dx^2}=\frac{1}{\Delta}\frac{1}{1+r}\gamma_x^1(x^2,k^1)c_{kx}^1(x_2^1,k^1)$$
$$\times\big[p''(x_1^1)x_1^1+2p'(x_1^1)-c_{xx}^1(x_1^1,k^1)\big]<0$$

Since $\gamma_k^1(x^2,k^1)>0$, the fact that the established firm's capital stock is lowered by a higher price of capital and by a higher interest rate implies that

$$(22)\qquad dx_2^1/dq<0$$

$$(23)\qquad dx_2^1/dr<0$$

In addition, equations (20) and (21) imply that the established firm chooses a *lower level of output and capital before entry* than a monopolist not anticipating entry. Clearly the firm lowers its output to accommodate entry. Equation (14) and $c_{kx}^1<0$ imply that

$$(24)\qquad -c_k(x_1^1,k^1)\left(1+\frac{1}{1+r}\right)>q$$

$$(25)\qquad -c_k(x_2^1,k^1)\left(1+\frac{1}{1+r}\right)<q$$

Thus, capacity k^1 is too low to produce x_1^1 efficiently in both periods and too high to produce x_2^1 efficiently in both periods.

IV. The Stackelberg Case

Suppose now that the established firm follows a Stackelberg leadership strategy in the post-entry game.[9] Then, given capacity k^1,

[9] Without explicitly allowing for excess capacity, Spence (1979) considers investment in a dynamic model with sequential entry and finds the dynamic Nash equi-

and the entrant's reaction function $x^2 = \gamma^2(x_2^1)$, the established firm solves the following problem in the second period:

$$(26) \quad \max_{x_2^1} \left[p\left(x_2^1 + \gamma^2(x_2^1)\right)x_2^1 - c^1(x_2^1)\right]$$

$$\text{subject to} \qquad x_2^1 \leqslant k^1$$

Let (S^1, S^2) be the *unconstrained* Stackelberg equilibrium. The equilibrium point occurs where the established firm's unconstrained iso-profit curve is tangent to the entrant's reaction curve (see Figure 4). When $k^1 \geqslant S^1$, then the established firm maximizes second-period profits by producing output at the Stackelberg equilibrium point. When $k^1 < S^1$, the established firm will choose $x_2^1 = k^1$ as this is on the lowest attainable iso-profit curve (see Figure 4). In this case, the Stackelberg equilibrium in the post-entry game is $(k^1, \gamma^2(k^1))$.

The value of post-entry capacity for the Stackelberg leader $V^S(k)$ is defined by:

$$(27) \quad V^S(k) = \begin{cases} p(S^1 + S^2)S^1 - c^1(S^1) \text{ if } k \geqslant S^1 \\ p\left(k + \gamma^2(k)\right)k - c^1(k) \text{ if } k \leqslant S^1 \end{cases}$$

It is assumed that the post-entry profit function $[p(x+\gamma^2(x))x - c^1(x)]$ is strictly concave. This is a sufficient condition for a unique maximum at the Stackelberg equilibrium. This assumption implies that $V^{S'}(k)$ is decreasing in k for $k \leqslant S^1$ and $V^{S'}(k) = 0$ for $k \geqslant S^1$. Using the implicit value of capacity, the established firm solves:

$$(28) \quad \max_{x,k} \left[\pi(x) - q \cdot k + \frac{1}{1+r} V^S(k) \right]$$

$$\text{subject to} \qquad x \leqslant k$$

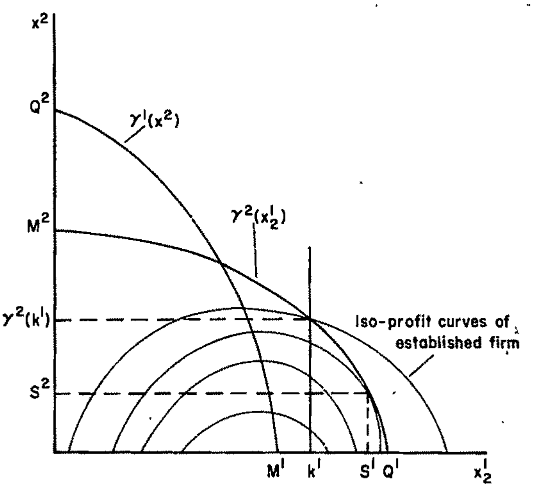


FIGURE 4. $k^1 < S^1$

The Kuhn-Tucker first-order necessary conditions for the problem are

$$(29) \qquad \pi'(x) = \theta$$

$$(30) \qquad \theta + \frac{1}{1+r} V^{S'}(k) = q$$

$$(31) \qquad \theta(k - x) = 0, \theta \geqslant 0$$

The conditions for entry deterrence when the established firm is a Stackelberg leader may be examined using (29)–(31). If the established firm's short-run monopoly output is greater than or equal to the entry blocking output $M^1 \geqslant Q^1$, the *unconstrained* Stackelberg leader's output is, of course, equal to M^1. As in the Cournot-Nash case, however, if capacity is too costly the established firm may still not find entry deterrence worthwhile. On the other hand, if $M^1 < Q^1$, the *unconstrained* Stackelberg output will not necessarily deter entry in the static second period game. Entry deterrence in the unconstrained static game only occurs, as is noted by Osborne (1973), when the entrant's reaction function is steeper than the (unconstrained) iso-profit contour of the established firm which passes through the blocking point Q^1. This slope condition is a necessary but not sufficient condition for the established firm to block entry when capacity is costly.

librium to be similar to the Stackelberg duopoly equilibrium. This occurs because the first firm to reach the Nash equilibrium capital stock may continue investing so as to reduce the follower's desired equilibrium capital stock, or even to deter the follower's entry. The model presented here clearly differs from Spence in that capacity investment is not in itself sufficient to deter entry.

This is especially true since the established firm must carry *unwanted excess capacity before entry.* Case (ii) of the following proposition establishes conditions for excess capacity before entry to be a desirable entry deterring strategy.

PROPOSITION 4: *If the established firm follows a Stackelberg strategy, then*

(32) (i) *For* $Q^1 \leqslant M^1$,

entry will be blocked if and only if

$$\pi'(Q^1)\left(1+\frac{1}{1+r}\right)+\frac{1}{1+r}\gamma^{2'}(Q^1)p'(Q^1)Q^1 \geqslant q$$

(33) (ii) *For* $Q^1 > M^1$,

entry will be blocked if and only if

$$\frac{1}{1+r}\left[\pi'(Q^1)+\gamma^{2'}(Q^1)p'(Q^1)Q^1\right]\geqslant q$$

Proposition 4 implies that *the established firm will hold excess capacity before entry and increase output to the entry-deterring level in the second period if* $Q^1 > M^1$ *and condition* (33) *is satisfied.* In addition, *the established firm will deter entry by holding output constant if* $Q^1 \leqslant M^1$ *and condition* (32) *is satisfied.* Thus, Proposition 4 implies that the behavior described by the Sylos Postulate and the Excess Capacity Hypothesis *exhausts the list of entry-deterring strategies.* As in the Cournot-Nash case, the pursuit of entry-deterring strategies is facilitated if the cost of capacity is low, the interest rate is low or the elasticity of demand at Q^1 is low.

Proposition 4 shows that even when the established firm behaves as a Stackelberg leader entry deterrence will not always be undertaken. The following results hold when entry is permitted as well as when entry is deterred.

PROPOSITION 5: *If the established firm follows a Stackelberg strategy, then*

(a) *If* $S^1 < M^1$, *then* (i) *The firm operates at full capacity before entry takes place. Capacity is strictly less than* M^1. (ii) *Capacity is greater than, equal to or less than* S^1 *if and only if* $\pi'(S^1)$ *is greater than, equal to or less than* q.

(b) *If* $S^1 = M^1$, *then the firm operates at full capacity before entry takes place. Capacity is strictly less than* $M^1 = S^1$.

(c) *If* $S^1 > M^1$, *then* (i) *The firm will hold excess capacity before entry takes place if and only if* $(1/1+r)V^{S'}(M^1)$ *is greater than* q. *In this case, output will equal* M^1 *in the first period.* (ii) *Capacity is greater than, equal to or less than* M^1 *if and only if* $(1/1+r)V^{S'}(M^1)$ *is greater than, equal to or less than* q. (iii) *Capacity is strictly less than* S^1.

PROOF:

(a) The argument is similar to the proof of Proposition 2.

(b) Suppose that $\theta=0$. This implies that the constraint $x \leqslant k$ is nonbinding and from (29), $x=M^1$. So $S^1 \leqslant k$. But this implies that $(1/1+r)V^{S'}(k)=0$ which contradicts (30). So $\theta>0$, and $x=k$. Since $\theta>0, x<M^1$ by (29), so $k<M^1$.

(c)(i) Suppose $x<k$. Then $\theta=0$ by (31) and $x=M^1$ by (29). Then by (30), $(1/1+r)V^{S'}(k)=q$, so $(1/1+r)V^{S'}(M^1)>q$. Conversely, suppose $(1/1+r)V^{S'}(M^1)>q$. Then by (30), $k>M^1$ and $x<k$. (ii) Suppose $k=M^1$. Then $x=k=M^1$ and $(1/1+r)V^{S'}(k)=q$. Suppose $k<M^1$. Then $x=k<M^1$ and since $\pi'(x)>0$, $(1/1+r)\ V^{S'}\ (k)<q$. So $(1/1+r)V^{S'}(M^1)<q$. Since this exhausts the possible cases the converse is also true. (iii) Suppose $k \geqslant S^1$. Then $(1/1+r)V^{S'}(k)=0$. So $\theta>0$, $x=k$ and $x<M^1$. So $k<M^1<S^1$. This is a contradiction, so $k<S^1$.

Thus we see from (c (i)) that *the established firm will only carry excess capacity before entry takes place if two conditions hold. First, the Stackelberg equilibrium output for the established firm must exceed its monopoly output. Secondly, the discounted marginal benefits of producing at the short-run monopoly level* M^1 *in the second period must exceed the cost of capacity.* In this case capacity k will satisfy:

$$\frac{1}{1+r}V^{S'}(k)=q \tag{34}$$

or, substituting for $V^{S'}(k)$,

$$(35)\qquad \frac{1}{1+r}\left[p'k+p+\gamma^{2'}(k)p'k\right]=q$$

where $p=p(k+\gamma^2(k))$. The established firm will produce at output level M^1 before entry occurs and then raise its output to k after entry has taken place.

In *all* cases *other* than (c (i)), *the established firm will produce at full capacity in the first period.* When $S^1<M^1$ and the marginal return to producing at the Stackelberg point does not exceed the cost of capital, the Sylos Postulate will be satisfied and the firm will produce at full capacity in both periods. When $S^1=M^1$, the Sylos Postulate is satisfied and the firm always operates at full capacity. Also, when $S^1>M^1$ and the discounted marginal benefits of producing at the monopoly level M^1 in the second period do not exceed the cost of capital, the established firm will operate at full capacity in each period, thus satisfying the Sylos Postulate. In these cases capacity will not exceed the Stackelberg output S^1 and satisfies equation (36):

$$(36)\qquad \pi'(k)+\frac{1}{1+r}V^{S'}(k)=q$$

The Sylos Postulate is violated, however, when the firm holds excess capacity in the first period (c (i)) or in the second period. The lowering of output below capacity occurs in the second period when $S^1<M^1$ and the marginal benefits from producing at level S^1 in the first period exceed the cost of capital.

V. The Stackelberg Case with the General Cost Function

Let the established firm have the general cost function introduced in Section III. Suppose that the entry blocking output Q^1 exceeds the established firm's monopoly output. Then, *the only active entry-deterring strategy will involve carrying excess capacity before entry for the purpose of expanding output in the post-entry game.* As was the case in the previous section, entry will be deterred and $x_2^1=Q^1$ if and only if the entrant's reaction function is steeper than the established firm's iso-profit contour in the second period. Thus, entry will be deterred in this case if there exists a $\tilde{k}$ large enough such that the slope conditions are satisfied which also solves

$$(37)\qquad p'(x_1^1)x_1^1+p(x_1^1)-c_x(x_1^1,\tilde{k})=0$$

$$(38)\qquad q-c_k(x_1^1,\tilde{k})-c_k(Q^1,\tilde{k})=0$$

Note that *pre-entry output and investment will be higher than under monopoly.*

Again, it must be emphasized that this special condition for entry deterrence is not satisfied in general. Without entry deterrence, a comparative static analysis of the post-entry outcome will yield results similar to Proposition 5. In particular, *pre-entry output and investment will be raised or lowered depending upon whether the Stackelberg leadership output in the post-entry game exceeds or is less than the established firm's monopoly output.* Further, *the established firm's output will rise or fall over time depending upon whether the Stackelberg leadership output exceeds or is less than the established firm's monopoly output.*

VI. Conclusion

An established firm faced with large-scale entry will only hold excess capacity before entry when the firm is a Stackelberg leader, the Stackelberg output exceeds the firm's short-run monopoly output and the discounted marginal value of producing at the pre-entry monopoly level in the second period exceeds the cost of capacity. An additional condition must be met for the firm to hold sufficient excess capacity for entry deterrence. These results imply that the Excess Capacity Hypothesis is only valid under quite limited conditions and that analyses employing it should certainly be avoided as guides to policy. A similar caveat applies to the Sylos Postulate. A constant output response to entry depends on the presence of costly and imperfectly adjustable inputs and entry deterrence depends on a rather special condition. This may have been Sylos-Labini's original intention when he wrote that established firms produce less than maximum output "not under pressure from new entry, but on the basis of independent economic calculations" (p. 43).

Becker and I perform an extensive analysis of the factor biases experienced by an established firm reacting to entry. There the post-entry cost function is derived through a cost-minimization approach when adjustment costs are present.

An additional extension of the present model is of interest. Since excess capacity observed in industry is in large part due to random demand fluctuations, the question of strategic capacity investment and entry needs to be reexamined when market uncertainty is present. The presence of risk will serve to emphasize strongly the advantage of flexibility which an entrant possesses as compared to an established firm.

REFERENCES

Joseph S. Bain, *Barriers to New Competition*, Cambridge 1956.

R. A. Becker and D. F. Spulber, "Entry, Strategic Investment and Factor Biases," work. paper no. 80-10, Brown Univ., July 1980.

J. N. Bhagwati, "Oligopoly Theory, Entry Prevention and Growth," *Oxford Econ. Papers*, Nov. 1970, *22*, 297–310.

Augustin Cournot, *Recherches sur les Principes Mathèmatiques de la Thèorie des Richesse*, Paris, 1838.

A. Dixit, "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell J. Econ.*, Spring 1979, *10*, 20–32.

______, "The Role of Investment in Entry-Deterrence," *Econ. J.*, Mar. 1980, *90*, 95–106.

F. M. Fisher, "New Developments on the Oligopoly Front," *J. Polit. Econ.*, Aug. 1959, *67*, 410–13.

C. R. Frank, "Entry in a Cournot Market," *Rev. Econ. Stud.*, July 1965, *32*, 245–50.

D. W. Gaskins, "Dynamic Limit Pricing: Optimal Pricing Under Threat of Entry," *J. Econ. Theory*, Sept. 1971, *3*, 306–22.

F. H. Hahn, "Excess Capacity and Imperfect Competition," *Oxford Econ. Papers*, Oct. 1955, *7*, 229–40.

R. F. Harrod, "Theory of Imperfect Competition Revised," in *Economic Essays*, New York 1952, 139–57.

J. R. Hicks, "The Process of Imperfect Competition," *Oxford Econ. Papers*, Feb. 1954, *6*, 41–54.

M. I. Kamien and N. L. Schwartz, "Uncertain Entry and Excess Capacity," *Amer. Econ. Rev.*, Dec. 1972, *62*, 918–27.

______ **and** ______, "Limit Pricing and Uncertain Entry," *Econometrica*, May 1971, *39*, 441–54.

M. McManus, "Equilibrium Numbers and Size in Cournot Oligopoly," *Yorkshire Bull. Econ. and Soc. Res.*, No. 2, 1964, *16*, 68–75.

F. Modigliani, "New Developments on the Oligopoly Front: Reply," *J. Polit. Econ.*, Aug. 1959, *67*, 418–19.

______, "New Developments on the Oligopoly Front," *J. Polit. Econ.*, June 1958, *66*, 215–32.

J. F. Nash, Jr., "Non-cooperative Games," *Annals of Mathematics*, 1951, *45*, 286–95.

Douglas Needham, *The Economics of Industrial Structure, Conduct and Performance*, New York 1978.

W. Novshek, "Cournot Equilibrium with Free Entry," *Rev. Econ. Stud.*, Apr. 1980, *47*, 473–86.

______ **and H. Sonnenschein,** "Cournot and Walras Equilibrium, *J. Econ. Theory*, Dec. 1978, *19*, 223–66.

K. Okuguchi, "Quasi-Competitiveness and Cournot Oligopoly," *Rev. Econ. Studies*, Jan. 1973, *40*, 145–48.

D. K. Osborne, "The Role of Entry in Oligopoly Theory," *J. Polit. Econ.*, Aug. 1964, *72*, 396–402.

______, "On the Rationality of Limit Pricing," *J. Ind. Econ.*, Sept. 1973, *22*, 71–80.

P. Pashigian, "Limit Price and the Market Share of the Leading Firm," *J. Indus. Econ.*, July 1968, *16*, 165–77.

R. Ruffin, "Cournot Oligopoly and Competitive Behavior," *Rev. Econ. Stud.*, Oct. 1971, *38*, 493–502.

F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

A. M. Spence, "Entry, Investment and Oligopolistic Pricing," *Bell J. Econ.*, Autumn, 1977, *8*, 534–44.

______, "Investment Strategy and Growth in a New Market, *Bell J. Econ.*, Spring 1979, *10*, 1–19.

Paolo Sylos-Labini, *Oligopoly and Technical Progress*, Cambridge, Mass. 1969.

J. T. Wenders, (1971a) "Excess Capacity as a Barrier to Entry," *J. Ind. Econ.*, Nov. 1971, *20*, 14–19.

______, (1971b) "Collusion and Entry," *J. Polit. Econ.*, Dec. 1971, *79*, 1258–77.

Multinational Firms and the Theory of International Trade and Investment: A Correction and A Stronger Conclusion

By A. WAHHAB KHANDKER*

In a recent article in this *Review*, Raveendra Batra and Rama Ramachandran showed that the traditional models of international trade can preserve most of the attributes introduced by multinational firms. As part of their exposition, they conducted a comparative statics analysis to explore the implications of taxes and tariffs on resource allocation and international capital movement. The purpose of this paper is to correct some of the comparative statics results. In Section I, I show some mathematical errors in Section IV of their paper. Making these corrections leads to some quantitative changes in their results even though their conclusions are still valid. In Section II, I will show that some stronger results follow so far as the implications of tariffs on the rental on multinational capital is concerned.

I. Implications of Tariffs on the Employment of Labor and Capital in Each Sector in Each Country

When $t=t^*=0$, equation (29) in their paper becomes

$$\begin{bmatrix} P^*X^*_{LL}+Y^*_{LL} & P^*X^*_{KL} & 0 \\ 0 & -PX_{KL} & PX_{LL}+Y_{LL} \\ P^*X^*_{KL} & P^*X^*_{KK}+PX_{KK} & -PX_{KL} \end{bmatrix}$$

$$\times \begin{bmatrix} dL^*_X \\ dK^*_X \\ dL_X \end{bmatrix} = -P \begin{bmatrix} X^*_L d\tau^* \\ 0 \\ X^*_K d\tau^* \end{bmatrix}$$

So, in equation (30) in their paper, G should be

$$G=-P\begin{bmatrix} X^*_L d\tau^* \\ 0 \\ X^*_K d\tau^* \end{bmatrix}$$

Correcting the calculations for the expressions $dL^*_X/d\tau^*$, $dK^*_X/d\tau^*$, and $dL_X/d\tau^*$ we get

$$\frac{dL^*_X}{d\tau^*}=$$

$$\frac{1}{D}\begin{vmatrix} -PX^*_L & P^*X^*_{KL} & 0 \\ 0 & -PX_{KL} & PX_{LL}+Y_{LL} \\ -PX^*_K & P^*X^*_{KK}+PX_{KK} & -PX_{KL} \end{vmatrix}$$

$$=\frac{1}{D}\{-PX^*_L[P^2X^2_{KL}-(P^*X^*_{KK}+PX_{KK})$$
$$(PX_{LL}+Y_{LL})]$$
$$-PX^*_KP^*X^*_{KL}(PX_{LL}+Y_{LL})\}$$

$$=\frac{1}{D}\{PX^*_L[P^*X^*_{KK}(PX_{LL}+Y_{LL})$$
$$+PX_{KK}Y_{LL}+P^2H_X]$$
$$-PP^*X^*_KX^*_{KL}(PX_{LL}+Y_{LL})\}$$

Hence the right-hand side of equation (31) in their paper not only needs to be multiplied by a constant P, but it lacks a P^* in front of X^*_K in the third line of their expression. By similar reasoning, it can be shown that

$$\frac{dK^*_X}{d\tau^*}=\frac{1}{D}\{P(PX_{LL}+Y_{LL})$$
$$\times[X^*_K(P^*X^*_{LL}+Y^*_{LL})-P^*X^*_LX^*_{KL}]\}$$

and $$\frac{dL_X}{d\tau^*}=\frac{1}{D}\{P^2X_{KL}[X^*_K(P^*X^*_{LL}+Y^*_{LL})$$
$$-P^*X^*_LX^*_{KL}]\}$$

*Assistant professor, department of economics, Wabash College.

Hence the right-hand side of equation (32) should be multiplied by P and it should be P^* instead of P^{*2} in the second line of that equation. Equation (33) of their paper is correct except that the right-hand side should be multiplied by P.

It is the signs and not the magnitudes of these expressions which determine their conclusions. In all the above cases, the signs of the corrected expressions are unaltered and hence there are no qualitative changes in their conclusions.

II. Effect of Tariffs on the Rental on the Multinational Capital

A more important change occurs when we derive the effect of a tariff on the rental on multinational capital. From $r_X = PX_K$ and $K_X + K_X^* = \overline{K}_X$,

$$\frac{dr_X}{d\tau^*} = P\left[\underset{(+)}{X_{KL}}\ \underset{(-)}{\frac{dL_X}{d\tau^*}} - \underset{(-)}{X_{KK}}\ \underset{(+)}{\frac{dK_X}{d\tau^*}}\right]$$

By observing the parallel movement of the two factors due to a tariff, Batra and Ramachandran concluded that a priori nothing can be said about the response of the rental on capital owned by multinational firms. However, substituting the corrected values of their equations (32) and (33), we get

$$\begin{aligned}\frac{dr_X}{d\tau^*} = \frac{P^3X_{KL}^2}{D}&\left[X_K^*(P^*X_{LL}^* + Y_{LL}^*)\right.\\ &\left.-P^*X_L^*X_{KL}^*\right] - \frac{P^2X_{KK}}{D}(PX_{LL} + Y_{LL})\\ &\times\left[X_K^*(P^*X_{LL}^* + Y_{LL}^*) - P^*X_L^*X_{KL}^*\right]\end{aligned}$$

$$\begin{aligned}= -\frac{P^2}{D}&\left[X_K^*(P^*X_{LL}^* + Y_{LL}^*) - P^*X_L^*X_{KL}^*\right]\\ &\times\left(PX_{KK}X_{LL} + X_{KK}Y_{LL} - PX_{KL}^2\right)\end{aligned}$$

$$\begin{aligned}= -\underset{(+)}{\frac{P^2}{D}}&\left[\underset{(+)}{X_K^*}\left(\underset{(+)}{P^*}\ \underset{(-)}{X_{LL}^*} + \underset{(-)}{Y_{LL}^*}\right)\right.\\ &\left.-\underset{(+)}{P^*}\ \underset{(+)}{X_L^*}\ \underset{(+)}{X_{KL}^*}\right]\left(\underset{(+)}{P}\underset{(+)}{H_X} + \underset{(-)}{X_{KK}}\ \underset{(-)}{Y_{LL}}\right)\end{aligned}$$

which is clearly positive. This implies that an imposition of tariffs on the host country's importables will definitely raise the rental on multinational capital. The economic explanation of this result is straightforward. At constant terms of trade, which is implied by the small country assumption, the imposition of tariffs on importables in the host country will increase the domestic price of importables by the amount of the tariff. This increases the marginal revenue product of multinational capital in the host country. The fixed supply of multinational capital implies that multinational capital moves from the source to the host country only at the cost of X production in the source country. As a result, the rental on multinational capital is bound to increase.

REFERENCE

R. N. Batra and R. Ramachandran, "Multinational Firms and the theory of International Trade and Investment," *Amer. Econ. Rev.*, June 1980, *70*, 278–90.

ERRATA

A Model of Sales

By HAL R. VARIAN*

The "density function" given after equation (15) in my article published in this *Review*, September 1980, is incorrect. This cannot possibly be a correct density function since it cannot integrate to one. I would like to thank Carl Norstrom of the Norwegian School of Economics and Business Administration for drawing this error to my attention. In fact, this expression gives the probability density that any *particular* store charges a price p and p is the lowest price charged. In order to compute the overall probability that p is the lowest price charged, we need to sum this probability over all stores. By virtue of symmetry:

$$f_{min}(p)=n(1-F(p))^{n-1}f(p)$$

It can easily be checked that the above formula for $f_{min}(p)$ does integrate to one.

This affects the subsequent analysis in the following ways:

1) Formula (16) now becomes:

$$\bar{p}_{min}=\frac{M}{I}(r-\bar{p})$$

2) In Table 1, the minus entry in the last row should be a question mark.

3) The effect described in the last paragraph before Section III *may* occur, but it does not *necessarily* occur. That is, more uninformed consumers *may* confer a beneficial externality on the informed consumers through an increase in the number of stores but this will not necessarily happen. Thus the paradoxical effect described in this part of the article is still present in the correct version of the model, but it is somewhat weakened.

4) The formula given in the text immediately before Section IV involves the expression for $\bar{p}_{min}$ and should be adjusted in accordance with the correction given above.

*University of Michigan.

Auditors' Report

To the Executive Committee of
The American Economic Association:

We have examined the statement of assets and liabilities of The American Economic Association (a District of Columbia corporation, not for profit) as of December 31, 1980 and 1979, and the related statements of revenues and expenses, changes in general and restricted fund balances and changes in assets and liabilities for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying financial statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1980 and 1979, and its revenues and expenses, changes in fund balances and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Arthur Andersen & Co.

Nashville, Tennessee,
February 26, 1981.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF ASSETS AND LIABILITIES
DECEMBER 31, 1980 AND 1979

Assets	1980	1979	**Liabilities and Fund Balances**	1980	1979
CASH	**$ 393,920**	**$ 121,183**	ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	**$ 214,073**	**$ 217,408**
			DEFERRED INCOME (Note 1):		
INVESTMENTS, at market (Notes 1 and 2):	**1,924,880**	**1,541,376**	Life membership dues	57,546	60,168
			Other membership dues	399,379	357,446
			Subscriptions	379,072	190,084
ACCOUNTS RECEIVABLE:			*Job Openings for Economists*	13,234	13,634
Advertising, back issues, etc.	96,891	108,859		**849,231**	**621,332**
Allowance for doubtful accounts	(1,519)	(404)	ACCRUAL FOR DIRECTORY (Note 1)	**139,102**	**80,689**
	95,372	**108,453**			
			FUND BALANCES:		
			Restricted (Note 4)	6,111	6,011
INVENTORY OF *Index of Economic Articles*, at cost	**55,557**	**37,268**	General	1,083,095	876,786
PREPAID EXPENSES	**22,014**	**17,427**	Add—Unrecognized change in market value of investments (Notes 1 and 3)	224,071	36,781
			General fund-net worth	1,307,166	913,567
			Total fund balances	1,089,206	882,797
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $11,723 in 1980 and $10,375 in 1979	**23,940**	**13,298**	Add—Unrecognized change in market value of investments (Notes 1 and 3)	224,071	36,781
			Net fund balance	**1,313,277**	**919,578**
Total Assets	**$2,515,683**	**$1,839,007**	**Total Liabilities and Fund Balances**	**$2,515,683**	**$1,839,007**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1980 AND 1979

	1980	1979
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 609,358	$ 554,502
Nonmember subscriptions	294,480	287,285
Job Openings for Economists subscriptions	20,109	19,281
Advertising	100,299	97,939
Sale of *Index of Economic Articles*	27,597	77,299
Sale of copies, republications, and handbooks	31,837	45,277
Sale of mailing list	33,397	34,246
Annual meeting	29,721	3,523
Sundry	20,598	21,673
	1,167,396	1,141,025
INVESTMENT GAINS (Note 2)	176,257	71,388
Net revenues	**1,343,653**	**1,212,413**
PUBLICATION EXPENSES:		
American Economic Review	398,721	314,635
Journal of Economic Literature	460,068	376,797
Directory publication (Note 1)	60,000	55,000
Job Openings for Economists	36,917	33,616
Index of Economic Articles	13,661	40,716
	969,367	820,764
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative—		
Salaries	115,131	107,635
Rent	10,414	10,137
Other (Exhibit I)	97,740	83,848
Committee	52,548	31,260
Annual meeting	4,715	3,889
Provision for federal income taxes (Note 6)	9,900	14,300
	290,448	251,069
Total expenses	**1,259,815**	**1,071,833**
REVENUES IN EXCESS OF EXPENSES	**$ 83,838**	**$ 140,580**

The accompanying notes to financial statements and Exhibit I are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN GENERAL FUND BALANCE FOR THE YEARS ENDED DECEMBER 31, 1980 AND 1979

	Total	Operations	Market Value Adjustments
Balance at December 31, 1978	**$674,494**	**$366,388**	**$308,106**
Add—market value adjustments resulting from inflation (Note 1)	61,712	–	61,712
Add—revenues in excess of expenses	140,580	140,580	–
Balance at December 31, 1979	**876,786**	**506,968**	**369,818**
Add—market value adjustments resulting from inflation (Note 1)	122,471	–	122,471
Add—revenues in excess of expenses	83,838	83,838	–
Balance at December 31, 1980	**$1,083,095**	**$590,806**	**$492,289**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1980

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains	Balance at December 31
Funds reserved by the Association for publication of revised editions of *Graduate Study In Economics*, a guide originally published with funds from a Ford Foundation grant	$ –	$4,182	$(4,182)	$ –	$ –
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	–	85,500	(85,500)	–	–
The Minority scholarship fund for minority students applying for graduate work in economics	5,000	–	–	–	5,000
Sundry	1,011	100	–	–	1,111
	$6,011	**$89,782**	**$(89,682)**	**$ –**	**$6,111**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1979

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics (Note 4)	$50,721	$ –	$ (60,195)	$9,474	$ –
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	–	67,335	(67,335)	–	–
Funds reserved by the Association for publication of revised editions of *Graduate Study In Economics*, a guide originally published with funds from a Ford Foundation grant	–	2,770	(2,770)	–	–
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	466	–	(466)	–	–
The Minority scholarship fund for minority students applying for graduate work in economics	5,000	–	–	–	5,000
The Ford Foundation grant for development of a consortium on graduate studies in economics for minorities	1,676	–	(1;676)	–	–
The International Communication Agency grant arranging for a delegation of American economists to participate in a joint symposium with the Soviet Federation of Economic Institutions	–	7,130	(7,130)	–	–
Sundry	911	100	–	–	1,011
	$58,774	**$77,335**	**$(139,572)**	**$9,474**	**$6,011**

The accompanying notes to financial statements are an integral part of this statement.

The American Economic Association Statement of Changes in Assets and Liabilities for the Years Ended December 31, 1980 and 1979

	1980	1979
Cash, beginning of year	**$121,183**	**$ 87,860**
Source (Use) of Funds:		
Revenues in excess of expenses	83,838	140,580
Add—noncash charges—		
Depreciation	1,836	1,295
Directory publication (Note 1)	60,000	55,000
Market value adjustments (Note 1)	(63,083)	11,430
Funds provided by operations	82,591	208,305
(Increase) decrease in—		
Receivables and prepaid expenses	8,496	742
Inventory of *Index of Economic Articles*	(18,289)	15,569
Investments	(383,504)	(208,191)
Office furniture and equipment	(12,478)	80
Increase (decrease) in—		
Accounts payable and accrued liabilities	(4,922)	(100,108)
Deferred income	227,899	38,827
Restricted funds	100	(52,763)
General fund, market value adjustment	122,471	61,712
Unrecognized change in market value of investments	250,373	69,150
Cash, end of year	**$393,920**	**$121,183**

The accompanying notes to financial statements are an integral part of this statement.

Notes to Financial Statements

(1) Significant Accounting Policies

Investments:

The Association accounts for its investments on a market value basis. According to the method the Association uses to value investments, the change in market value of corporate stock, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (10% in 1980 and 9.0% in 1979), is recognized in income over a three-year period for corporate stock and reflected in current income for government obligations, bonds and commercial paper. The changes in market value of investments are allocated to the general and restricted fund balances as appropriate.

Accrual for Directory:

Approximately every three to five years, the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was published in 1978 and distributed at no cost to the membership. In order to match more properly the publishing cost of this directory with revenue from membership dues, the Association provided $60,000 in 1980 and $55,000 in 1979 for estimated publishing costs which will reduce actual directory expense in the year of publication.

Deferred Income:

Revenue from membership dues and subscriptions to the various periodicals of the Association are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association.

Revenue from life membership dues is recognized over the estimated average life of these members.

(2) Investments and Investment Income

The following is a summary of investments held by the Association at December 31:

	1980		1979	
	Cost	Market	Cost	Market
Government obligations, bonds, and commercial paper	$ 447,828	$ 437,861	$ 602,106	$ 602,106
Corporate stocks	825,191	1,487,019	678,589	939,270
	$1,273,019	**$1,924,880**	**$1,280,695**	**$1,541,376**

Investment gains (losses) recognized in income for the years ended December 31, 1980 and 1979 were as follows:

	1980	1979
Government obligations, bonds, and commercial paper—		
Interest	$69,288	$50,722
Increase (decline) in market value recognized	(47,521)	–
	21,767	**50,722**
Corporate stocks—		
Cash dividends	43,886	32,096
Increase (decline) in market value recognized (Note 3)	110,604	(4,155)
	154,490	**27,941**
Less Investment gains allocated to restricted fund (Note 4)	–	7,275
Investment gains included in income	**$176,257**	**$71,388**

(3) Unrecognized Change in Market Value of Investments

As described more fully in Note 1, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1980 and 1979 revenues, and the amount of these market value increases (declines) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change December 31	
	1980	1979	1981	1982	1980	1979
1977	$ –	$(15,461)	$ –	$ –	$ –	$ –
1978	(14,168)	(14,169)	–	–	–	(14,168)
1979	25,474	25,475	25,475	–	25,475	50,949
1980	99,298	–	99,298	99,298	198,596	–
	$110,604	**$ (4,155)**	**$124,773**	**$99,298**	**$224,071**	**$36,781**

(4) Restricted Fund

The Economics Institute is an organization that provides orientation programs for foreign graduate students of economics. The Policy and Advisory Board which determines overall policies applicable to The Economics Institute is appointed by the President of the Association. In October 1979, The Economics Institute incorporated as an entity distinct from the Association. Prior to October 1979, The Economics Institute participated in the investment program of the Association and its share of investments were accounted for as restricted funds in the accompanying statement of assets and liabilities. In October 1979, The Economics Institute liquidated its share of the Association's investments. Investment income and market value adjustments applicable to The Economics Institute which were allocated to the restricted fund prior to liquidation were as follows:

	1979
Net investment gains (Note 2)	$7,275
Market value adjustments arising from inflation	2,199
	$9,474

(5) Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was $18,810 and $16,436 for 1980 and 1979, respectively.

(6) The Association

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c)(3) of the U.S. Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1980 AND 1979

	1980	1979
Mailing list file maintenance and periodic mailing expenses	$30,762	$28,201
Accounting and legal	10,800	9,600
Office supplies	12,601	7,659
Postage	12,660	14,210
Dues and subscription	5,390	5,895
Telephone	2,654	2,543
Investment counsel and custodian fees	6,969	4,499
President and president-elect expenses	4,786	3,985
Travel and entertainment	539	726
Depreciation (straight-line method)	1,836	1,295
Uncollectible receivables	1,134	507
Currency exchange charges	572	1,214
Insurance and miscellaneous	7,037	3,514
	$97,740	$83,848

NOTES

Nominations for AEA Officers: 1982

The Electoral College on March 21 chose W. Arthur Lewis as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1981. Other nominees (chosen by the 1981 Nominating Committee) are: Vice President (two to be elected), Robert Dorfman, Alfred E. Kahn, Allan H. Meltzer, and Edwin S. Mills; for members of the Executive Committee (two to be elected), Peter A. Diamond, Richard A. Easterlin, Joseph E. Stiglitz, and Ann F. Friedlaender.

Under a change in the bylaws as described in the *Papers and Proceedings* of this Review, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

1982 Nominating Committee of AEA

In Accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Gardner Ackley has appointed a Nominating Committee for 1982 consisting of Robert M. Solow, Chair; Marcus Alexius, Charles C. Holt, Daniel J. B. Mitchell, William D. Nordhaus, Steven Salop, and Isabel Sawhill. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

Program plans for the National Bureau of Economic Research Conference on Econometrics and Mathematical Economics (CEME), sponsored by the National Science Foundation, are being formulated for 1981-82. Suggestions for new topics and initiatives are being sought for consideration by the CEME steering committee.

Recommendations to the CEME steering committee for topic fields, particular seminars, subjects, papers, and participation of scholars, should be sent to the CEME Executive Secretary, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138.

On November 20-21, 1981, the National Bureau of Economic Research will hold a conference in Cambridge on Exchange Rates and International Macroeconomics. The program, being organized by Professor Jacob A. Frenkel of the University of Chicago and the NBER, will consist of seven papers with two formal discussants assigned to each paper. There will be no published proceedings, but papers will be included in the NBER Conference Paper series and summarized in the NBER *Reporter* and a special "Conference Summary Report." The Conference will be broad enough to accommodate a wide variety of issues relating to international macroeconomics. Priority will be given to empirically oriented research, but submission of theoretical papers is also welcomed. Papers will be selected on the basis of abstracts of about 500 words or, when possible, completed papers, with preference being given to papers by younger members of the profession. Any research that will not have been published at the time of the conference may be submitted. The deadline for submissions of abstracts and papers is June 20, 1981. Authors chosen to present papers will be notified by July 18. Finished papers must be ready for distribution to conference participants by October 16, 1981. The NBER will pay the expenses of those chosen to give papers at the conference. Abstracts and papers should be sent to Professor Jacob A. Frenkel, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138.

The Fifth International Time Series Meeting is being held in Houston, Texas, August 6-7, 1981. Interested persons are invited to register. Send all inquiries to M. Ray Perryman, Director, Center for the Advancement of Economic Analysis, Baylor University, Waco, TX 76798.

The Second Conference on Ukrainian Economics, "Selected Contributions of Ukrainian Scholars to Economics and Related Sciences" will be held at the Harvard Ukrainian Research Institute, 1581-83 Massachusetts Avenue, Cambridge, MA 02138, September 25-26, 1981. The conference will consist of two sessions at which papers on the work of M. Tuhan-Baranovsky, E. Slutsky, P. Liaschenko, M. Ptukha, R. Rosdolsky, and V. Holubnychy as well as on the current research on Mathematical Economics in Kiev will be presented and discussed. For further information, contact Professor I. S. Koropeckyj, Department of Economics, Temple University, Philadelphia, PA 19122.

The Fifth National Forum on Jobs, Money, and People: "Financing the Future: Political and Economic Incentives," will be held October 12-14, 1981, at the Shoreham Hotel, Washington, D.C. The cosponsors are the Fiscal Policy Council and the Chair of Private Enterprise, Georgia State University.

The First Conference on Regional Impacts of Mexico-United States Economic Relations will be held in Guanajuato, Mexico, July 8-11, 1981. The program will feature both invited and contributed papers which make a contribution to the north-south dialogue. Major themes are: Structure of the Mexican and American Economies; International Balance of Payments; Migration, Trade, and Capital Transfers between the United States and Mexico; Interregional Disparities—Mexico and the United States; and Development of the Border. Papers will be presented in either Spanish or English, and abstracts will be published in both languages. For program information, contact Lay James Gibson, Department of Geography and Regional Development, College of Business and Public Administration, University of Arizona, Tucson, AZ 85721.

International Conference on Time Series Methods in the Hydrosciences to be held October 6-8, 1981, Burlington, Ontario; organized by National Water Research Institute and cosponsored by Research and Development Division, Ocean and Aquatic Sciences (both of Canada Centre for Inland Waters) and Water Resources Branch, Ontario Ministry of Environment. Cochairmen are Oliver D. Anderson and Abdel H. El-Shaarawi. Send inquiries to Dr. El-Shaarawi, Aquatic Physics and Systems Division, NWRI, Canada Centre for Inland Waters, PO Box 5050, Burlington L7R 4A6, Ontario, Canada (telephone 416+637-4584).

A special summer program, "Forecasting Transportation Demand," will be held at Massachusetts Institute of Technology, August 3-21, 1981. The course consists of three parts which can be taken individually: Basic Concepts (Aug. 3-7), Modeling Techniques (Aug. 10-14), and Data Collection and Advanced Modeling Methods (Aug. 17-21). For further information, contact Director of Summer Session, Rm E19-356, MIT, Cambridge, MA 02139.

New Journal: *Marketing Science*, sponsored jointly by The Institute of Management Sciences and Operations Research Society of America, invites quantitatively oriented marketing manuscripts that make a significant contribution to the understanding of marketing phenomena and/or improve the practice of marketing management. The style format of *Management Science* should be followed with one exception: Bibliography and text reference should be in the style of *JM/JMR*. Submit manuscripts (four copies) to Professor Donald G. Morrison, Graduate School of Business, 414 Uris Hall, Columbia University, New York,. NY 10027. The first issue is planned for January 1982.

The *Monograph Series in Finance and Economics* is published by the Salomon Brothers Center for the Study of Financial Institutions, New York University Graduate School of Business Administration. The *Monograph Series* publishes medium length manuscripts that are too long and/or too broad in scope for scholarly journals or business periodicals. Manuscript submissions are welcomed and should be sent (in duplicate) to the editors: Professors Ernest Bloch and Lawrence J. White, Graduate School of Business Administration, New York University, 90 Trinity Place, New York, NY 10006.

Users of the Retirement History Study: A list of finished research projects which have used the Social Security Administration's Retirement History Study is being compiled. Because of privacy laws, persons who purchased the data tapes from the National Archives are not known. Also, many researchers have acquired tapes from other sources (Duke, Michigan, other users, etc.). Researchers inside and outside the government have expressed interest in exchanging findings from the Retirement History Study. If you have completed work (published or unpublished), please send a copy to the address given. If you would like to receive the list of finished analyses from the study, contact Dr. Lola Irelan, Director of the Retirement History Study, Room 1118-C, 1875 Connecticut Ave., NW, Washington, D.C. 20009.

College Curriculum Support Project (CCSP) *Update #3* is now available from the Bureau of the Census. It has two major sections: "Curriculum Development within the Bureau of the Census," and "1980 Census of Population and Housing." An extensive collection of 1980 census-related resources is also provided in the form of a three-part bibliography. *CCSP Update* will be published three times a year to highlight issues, products, and services of interest to the academic community. Persons who are on the CCSP mailing list automatically receive each *Update*. To be placed on this list and to receive a sampler of CCSP instructional materials, contact Les Solomon, CCSP, Data User Services Division, Bureau of the Census, Washington D.C. 20233 (telephone 301+449-1655).

The Law and Economics Center of the University of Miami, Coral Gables, presented its $1,000 Prize for Distinguished Scholarship in Law and Economics for

1979-80 to Ronald H. Coase, Professor of Economics, University of Chicago Law School.

Fulbright Fellowships: Academic institutions in the *USSR* wish to sponsor American lecturers, in any field, who can teach in languages of the constitutent republics. Nominations for 1981-82 have already been made, but scholars who wish to teach in the Soviet Union in 1982-83 and who are proficient in one of the following languages are invited to express that interest at an early date: Armeninan, Azerbaijani, Estonian, Georgian, Latvian, Lithuanian, Russian, or any of the Central Asian languages. For further information, please contact W. A. James, Council for International Exchange of Scholars, 11 Dupont Circle, NW, Washington, D.C. 20036 (telephone 202+833-4990).

More than 500 Fulbright awards in over 100 countries are now open to application for university teaching and postdoctoral research in 1982–83. Deadlines are June 1, 1981 for the American Republics, Australia, and New Zealand; and July 1, 1981 for Africa, Asia, Europe, and the Middle East. An applicant must be a U.S. citizen and have appropriate academic and experience credentials. Eligible scholars not available for 1982–83, but interested in a later possibility, may qualify to receive major announcements for the next two years by completing a registration form available from the Council for International Exchange of Scholars, Dept. *N*, Eleven Dupont Circle, NW, Suite 300, Washington, D.C. 20036.

The Bank of Israel has established the David Horowitz Memorial Prize, in honor of the founder and first Governor of the Bank of Israel. The prize of $5,000 (US) will be awarded to an outstanding work, theoretical or empirical, on a subject related to the role of central banks in the domestic and international economy. Works submitted shall be previously unpublished and shall be in the English or Hebrew languages. Closing date for entries is March 1, 1982 and the prize will be awarded in August 1982. Further information can be obtained from The Secretary, The Horowitz Prize, P.O.B. 780, Jerusalem, Israel 91000.

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics announce the annual Leonard J. Savage Award of five hundred dollars ($500) for an outstanding doctoral dissertation in the area of Bayesian Econometrics and Statistics. To be considered for the 1981 Savage Award, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1981 and accompanied by a short letter from the supervisor summarizing the main results of the dissertation. Dissertations completed after January 1, 1977 are eligible to be considered for the 1981 Savage Award. An Evaluation Committee will be appointed by the board of the Leonard J. Savage Memorial Trust Fund (M. H. DeGroot, S. E. Fienberg, S. Geisser, J. B. Kadane, E. E. Leamer, J. W. Pratt, and A. Zellner, chairman) to evaluate dissertations that are submitted for the Savage Award. Dissertations and supporting letters should be sent to Professor Arnold Zellner, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637.

The 1980 winner of the Award was Paul R. Milgrom. His doctoral dissertation "The Structure of Information in Competitive Bidding" was completed at Stanford University. The Thesis Evaluation Committee included Edward E. Leamer, Chairman, Christopher B. Barry, Nicholas M. Kiefer, Richard E. Kihlstrom, and Arnold Zellner (*ex officio*).

The Population Council, with funding from the United States Agency for International Development, has organized an International Research Awards Program on the Determinants of Fertility in Developing Countries. The objective of the program is to support research that will increase understanding of why and how human fertility changes in different cultural settings and under varying socioeconomic conditions. Inquiries about the program and application procedures are available from the Program Manager, Charles Keely, The Population Council, One Dag Hammarskjold Plaza, New York, NY 10017.

The Graduate School of Industrial Administration, Carnegie-Mellon University, announces a postdoctoral fellowship program in political economy. The GSIA fellowship provides a unique opportunity for a researcher with a strong commitment to the use of mathematical or quantitative analysis in the study of politics and the interdependence of political and economic decision making. Fellows are to devote a twelve-month period of residence to research. There are no teaching duties. Fellows may also take advantage of GSIA's doctoral program to obtain additional training in advanced topics. Applicants may either forward a resume and a brief statement of research interests or write for further information to Professor Howard Rosenthal, GSIA, Carnegie-Mellon University, Pittsburgh, PA 15213.

The School of Urban and Public Affairs at Carnegie-Mellon University, with the support of a training grant from the Center for Studies in Crime and Delinquency of the National Institute of Mental Health, is offering a postdoctoral program in Quantitative Methods in Criminal Justice to bring together specialists in disciplines related to the problems of crime with persons whose principal training is in methodology. In addition to a stipend of at least $13,380 (depending on years of postdoctoral experience), all training costs and research resources are provided by the training grant. A limited number of *pre*doctoral fellowships are also available to individuals who have already completed at least two years of graduate study. Participation in the program can begin in July 1981. Applications should be submitted as early as possible before that date. For further information, and for application forms, contact Professor Alfred Blumstein, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA 15213.

Call for Papers: The third annual North American meeting of the International Association of Energy Economists will be held November 12-13, 1981. The conference will focus on the following broad theme areas: Regulatory Issues for the 1980's; Energy/Environment in the 1980's; the Problems of Energy Disruptions; the Outlook for Domestic Oil and Natural Gas; International Trade and Energy, Development and Diffusion of New Technologies; Conservation and Energy Demand Analysis, Energy, and the Developing World. The conference will include both invited papers and contributed papers. Papers within the theme areas are preferred, although those not falling within these themes might be accepted. Authors wishing to present papers should submit abstracts to the Program Chairman immediately. Completed papers will be required by August 15. Abstracts should be sent to Professor James L. Sweeney, Engineering-Economic Systems Department, Terman 406, Stanford University, Stanford, CA 94305. Other inquiries about the conference should be addressed to William Hughes, Charles River Associates, John Hancock Tower, 200 Clarendon, Boston, MA 02116.

Call for Papers: Conference to be held March 4-6, 1982 will commemorate the centennial of Franklin Delano Roosevelt's birth and the fiftieth anniversary of his election as President of the United States. Papers dealing with his life, career, and presidency are invited. Because of her unique contribution, papers pertinent to Eleanor Roosevelt (1884-1962) are also invited. Deadline for completed papers is November 1, 1981, and must be submitted in duplicate. Contact Natalie Datlof and Alexej Ugrinsky, Conference Coordinators, University Center for Cultural and Intercultural Studies, Hofstra University, Hempstead, New York 11550 (telephone 516+560-3296, 3513, 3514).

Call for Papers: The Fourth Annual Conference on Public Finance will be held Thursday, July 2, 1981 at the Hyatt Regency Hotel, Embarcadero Center, San Francisco. (This is the day the Western Economics Association Meetings begin at the same location.) Abstracts of proposed papers, as well as other correspondence, should be sent to the program chairman, Professor W. Craig Stubblebine, Vice President, Western Tax Association, Department of Economics, Bauer Center, Claremont Men's College, Claremont, CA 91711 (telephone 714+621-8012).

Call for Papers: An international conference, "Urban and Regional Change in the Developing Countries," will be held at the Indian Institute of Technology at Kharagpur (Calcutta), India, December 13-17, 1981. This conference will cover such topics as rural-urban migration, housing, transportation, land use, economic development, population growth, urban/regional conflicts, etc. Papers related to developed countries, but having some significance for the developing countries are also welcome. Persons interested in presenting a paper should contact Professor Manas Chatterji, School of Management, State University of New York, Binghamton, NY 13901.

Call for Papers: The Rocky Mountain Conference on British Studies will hold its annual meeting on November 5-7, 1981, at the University of Nevada, Reno. Paper proposals should be submitted by July 15, 1981 to F. Darrell Munsell, Department of History, West Texas State University, Canyon, TX 79016. Further information about local arrangements is available from Neal Furguson, Continuing Education: College Inn, University of Nevada, Reno, NV 89557.

Call for Papers: The annual meeting of the Association of Environmental and Resource Economists (AERE) will be held jointly with the AEA in Washington, D.C., Dec. 28-30, 1981. One session of the meetings will be devoted to contributed papers from AERE members. Send two copies of a one-page abstract by July 15, 1981 to either Professor Geoffrey Heal (President-Elect), Department of Economics, University of Essex, Wivenhoe Park, Colchester C04 3SQ, Essex, England, or Professor V. Kerry Smith, Department of Economics, University of North Carolina, Chapel Hills, NC 27514.

Call for Papers: Aussenwirthschaft—The Swiss Review of International Economic Relations invites papers for a special issue. The topic is International Debt: The Long-Term Outlook. Articles (maximum length 30 pages double spaced) should be submitted no later than July 15, 1981, to *Aussenwirthschaft*, Schweizerisches Institut für Aussenwirthschaft, Dufourstrasse 48, CH-9000 St. Gallen, Switzerland. Selection will be on the basis of originality of approach and feasibility of measures proposed.

Call for Papers: The Academy of Criminal Justice Sciences is soliciting abstracts from persons interested in participating in the 1982 Annual Meeting, March 23-27, Louisville, KY. The Meeting Theme is "Interdisciplinary Contributions to Criminal Justice." For an abstract format, contact Robert G. Culbertson, President, ACJS, 401 SH, Illinois State University, Normal, IL 61761.

Call for Papers: In conjunction with its research program on youth employability, the National Center for Research in Vocational Education (NCRVE) announces its second annual *National Leadership for Research* competition. The 1981 competition seeks to encourage the submission of scholarly papers on issues relating to *Research and Policy Impacting on Youth Employability*. Papers can be submitted by either advanced graduate student/faculty advisor teams or by advanced graduate students accompanied by written letter of recommendation from their faculty advisor. Graduate students from the following academic disciplines are invited to submit: Psychology, Sociology, Economics, Anthropology, Education, Social Work, Business, or Labor and Human Resources. All papers will be reviewed both by NCRVE staff and an editorial review board of leading researchers from the aforementioned disciplines. Four or five winning papers will be selected from among those reviewed to be published as a monograph by the NCRVE. The finalists will also receive an honorarium of $400–$500. The deadline for submission is August 1, 1981. For information on eligibility and submission requirements contact Mrs. Barbara Fleming, Competition Coordinator, NCRVE, The Ohio State University, 1960 Kenny Road, Columbus, OH 43210.

1982-83 Advanced Research Fellowships in India: Twelve long-term (6-9 months) and nine short-term (2-3 months) awards, without restriction to field, are offered by the Indo–U.S. Subcommission on Education and Culture. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. Fellowship terms include $1,000–1,500 per month, depending on academic/professional achievement and seniority, $350 per month in dollars, balance in rupees; an allowance for books and study/travel in India; and international travel for grantee. In addition, long-term fellows receive international travel for dependents; a dependent allowance of $100–250 per month in rupees; and a supplementary research allowance up to 34,000 rupees. The application deadline is July 1, 1981. Application forms and further information are available from the Council for International Exchange of Scholars, ATT: Indo-American Fellowship Program, Eleven Dupont Circle, Washington, D.C. (telephone 202+833-4978).

1982 Near East/South Asia Short-Term Lectureships: Ten awards of six weeks to four months, beginning February 1982 and September 1982, are offered without restrictions as to field. Applicants must be U.S. citizens and have postdoctoral or equivalent professional level. Grant terms include roundtrip economy-class air travel and a per diem of $75 plus a temporary living allowance, not to exceed $125 per day. Application deadline is July 1, 1981. Additional information and applications are available from the Council for International Exchange of Scholars, ATT: Near East/South Asia Short-Term Program, Eleven Dupont Circle, Washington, D.C. (telephone 202+833-4981).

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. to be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Russell C. Hill, University of South Carolina.

Elizabeth E. Hoyt, professor of economics, Iowa State University, Nov. 22, 1980.

Benjamin A. Rogge, distinguished professor of political economy, Wabash College, Nov. 17, 1980.

Retirements

Robert W. Graham, Jr., University of South Carolina, fall 1980.

Promotions

M. Akbar Akhtar: manager, international research department, Federal Reserve Bank of New York, Jan. 1, 1981.

Paul B. Bennett: chief, Business Conditions Division, domestic research department, Federal Reserve Bank of New York, Jan. 1, 1981.

John H. Bradley: professor of economics, Biscayne College, Jan. 1, 1981.

Robert A. Brusca: chief, International Financial Markets Division, financial markets department, Federal Reserve Bank of New York, Jan. 1, 1981.

William A. Darity, Jr.: associate professor, University of Texas-Austin, Sept. 1, 1981.

Robert Foster: associate professor of economics, American Graduate School of International Management, Jan. 24, 1980.

Edward J. Frydl: manager, financial markets department, Federal Reserve Bank of New York, Jan. 1, 1981.

Alfred Hagan: associate professor of marketing, American Graduate School of International Management, Aug. 21, 1980.

Theresa Hagan: assistant professor of accounting, American Graduate School of International Management, Aug. 21, 1980.

F. Gregory Hayden: professor of economics, University of Nebraska-Lincoln, Aug. 1980.

David T. King: chief, Industrial Economies Division, international research department, Federal Reserve Bank of New York, Jan. 1, 1981.

Randolph D. Martin: professor, department of economics, University of South Carolina.

Ann Marie Meulendyke: research officer and senior economist, Open Market Operations and Treasury Issues Function, Federal Reserve Bank of New York, Jan. 1, 1981.

John Partlan: senior economist, Monetary Analysis Division, monetary analysis department, Federal Reserve Bank of New York, Jan. 1, 1981.

Leonard G. Sahling: manager, domestic research department, Federal Reserve Bank of New York, Jan. 1, 1981.

Don L. Schmidt: assistant professor of management, American Graduate School of International Management, Aug. 21, 1980.

Bernard L. Weinstein: professor of economics and political economy, University of Texas-Dallas, Sept. 1, 1980.

Steven R. Weisbrod: chief, Banking Studies Division, banking studies department, Federal Reserve Bank of New York, Sept. 11, 1980.

Ronald P. Wilder: professor, department of economics, University of South Carolina.

Roger J. Williams: professor, department of economics and finance, St. John's University, Feb. 27, 1980.

Mark A. Willis: senior economist, Regional Economics Staff, domestic research department, Federal Reserve Bank of New York, Jan. 29, 1981.

Administrative Appointments

John H. Bradley: chairman, Division of Transportation, Travel, and Tourism; director of business programs, Center for Continuing Education, Biscayne College, Jan. 1, 1981.

Thomas M. Freeman, Michigan State University: associate vice chancellor, Central Administration, State University of New York, Aug. 1, 1980.

Jess P. Hewitt: administrative analyst, Graduate Training Program, Transco Companies, Inc.

Marcos T. Jones, chief, Domestic Financial Markets Division, financial markets department, Federal Reserve Bank of New York, Jan. 1, 1981.

Craig R. MacPhee: chair, department of economics, University of Nebraska-Lincoln, Aug. 1980.

Wolfgang Mayer: director of graduate studies, department of economics, University of Cincinnati, Oct. 1980.

Roger LeRoy Miller: associate director, Law and Economics Center, University of Miami, Jan. 1981.

William H. Shaw: director, Council of Professional Associations on Federal Statistics, Washington, D.C., Dec. 1980.

Lloyd M. Valentine: head, department of economics, University of Cincinnati, Sept. 1980.

Betsy B. White: chief, Monetary Analysis Division, monetary research department, Federal Reserve Bank of New York, Jan. 1, 1981.

Appointments

John T. Addison: associate professor, department of economics, University of South Carolina.

Peter F. Allgeier, Agency for International Development: economist, Office of the U.S. Trade Representative, June 1980.

James Robert Alm, University of Wisconsin: assistant professor, department of economics, Syracuse University, Sept. 1, 1980.

Gary Brown: associate professor, department of economics, University of Cincinnati, Sept. 1980.

James R. Capra: senior economist, Fiscal Analysis Staff, monetary research department, Federal Reserve Bank of New York, Jan. 1, 1981.

Daniel E. Chall: economist, Business Conditions Division, domestic research department, Federal Reserve Bank of New York, Dec. 8, 1980.

Henry W. Chappell, Jr.: assistant professor, department of economics, University of South Carolina.

C. Lon Chen: assistant professor of economics, Whitman College, Sept. 1, 1980.

Linda R. Cohen, Harvard University: associate economist, economics department, The Rand Corporation, June 1980.

James N. Devine: assistant professor, Occidental College, Sept. 1980.

Sanjay Dhar: economist, Developing Economies Division, external financing department, Federal Reserve Bank of New York, July 30, 1980.

Michael Dotsey: economist, Monetary Analysis Division, domestic research department, Federal Reserve Bank of New York, Sept. 22, 1980.

Alfred S. Englander: economist, Business Conditions Division, domestic research department, Federal Reserve Bank of New York, Oct. 20, 1980.

Howard Y. Esaki: economist, Domestic Financial Markets Division, financial markets department, Federal Reserve Bank of New York, Nov. 5, 1980.

James R. Follain, Jr., Federal Home Loan Bank of San Francisco: assistant professor, department of economics, Syracuse University, Sept. 1, 1980.

Richard D. Gustely: director of business and economic research, and professor of economics, Oklahoma State University, Aug. 15, 1980.

Michael R. Haines, Cornell University: associate professor, department of economics, Wayne State University, Sept. 1980.

Werner Hochwald, Washington University: professor of economics, The University of the South, Spring 1981.

Shafigul Islam: economist, Industrial Economies Division, industrial research department, Federal Reserve Bank of New York, Dec. 22, 1980.

Daniel F. Kohler, Wayne State University: associate economist, economics department, The Rand Corporation, May 1980.

Lawrence Kreicher: economist, International Financial Markets Division, international research department, Federal Reserve Bank of New York, Sept. 3, 1980.

Ronald Levin: economist, Developing Economies Division, Federal Reserve Bank of New York, Oct. 20, 1980.

Edward B. Leviton, director of research, National Automobile Dealers Association, Nov. 1980.

Timothy Lord: economist, Financial Markets Division, domestic research department, Federal Reserve Bank of New York, July 28, 1980.

Robert G. McGillivray: senior economist, Applied Economics Associates, Inc., Seattle, Oct. 1, 1980.

Clair McRostie: visiting professor of economics, American Graduate School of International Management, Aug. 20, 1980.

Sholeh Maani: assistant professor, Occidental College, Sept. 1980.

Susan Foster Moore: chief, Market Reports Division, statistics department, Federal Reserve Bank of New York, Jan. 1, 1981.

Vincent G. Munley, President's Council on Wage and Price Stability: assistant professor, department of economics, Lehigh University, Aug. 26, 1980.

Michael P. Murray, Duke University: economist, The Rand Corporation, Aug. 1980.

Samuel L. Myers, Jr., University of Texas: staff economist, Federal Trade Commission, Jan. 1981.

Barbara Nunemaker: economist, Development Economies Division, external financing department, Federal Reserve Bank of New York, July 30, 1980.

William Petersen: economist, Banking Studies Division, banking studies department, Federal Reserve Bank of New York, Sept. 22, 1980.

William H. Phillips: assistant professor, department of economics, University of South Carolina.

Frank J. P. Pinto: consultant to the Assistant-Secretary-General for Development Research and Policy Analysis, department of international economic and social affairs, United Nations, Oct. 1980.

Allen Proctor: economist, Industrial Economies Division, international research department, Federal Reserve Bank of New York, Aug. 18, 1980.

Lawrence Radeki: economist, Monetary Analysis Division, domestic research department, Federal Reserve Bank of New York, July 28, 1980.

Rosemary Rossiter, University of Wisconsin: assistant professor, department of economics, Wayne State University, Sept. 1980.

Carlos Santiago, Cornell University: assistant professor, department of economics, Wayne State University, Sept. 1980.

Fredericka Santos: economist, Banking Studies Division, banking studies department, Federal Reserve Bank of New York, Sept. 3, 1980.

Allen Scafuri, Bowling Green State University: lecturer, department of economics, Wayne State University, Sept. 1980.

Joseph Scherer, professor of finance, department of banking and finance, School of Business, Hofstra University, Sept. 1980.

Sherrill Shaffer: economist, Banking Studies Division, banking studies department, Federal Reserve Bank of New York, Sept. 10, 1980.

Andrew Silver: economist, Financial Markets Division, domestic research department, Federal Reserve Bank of New York, Oct. 8, 1980.

Michael Thomson, Michigan State University: assistant professor, department of economics, Wayne State University, Sept. 1980.

Wen-Lee Ting: assistant professor of marketing, American Graduate School of International Management, Aug. 20, 1980.

Socrates Tountas, University of Michigan: lecturer, department of economics, Wayne State University, Sept. 1980.

Michael P. Ward, University of California-Los Angeles: economist, The Rand Corporation, June 1980.

Joseph A. Whitt, Jr.: assistant professor, department of economics, University of South Carolina.

Leaves for Special Appointment

C.-René Dominique, Laval University, Quebec: Harvard Institute for International Development, 1981-82.

Robert T. Falconer: consultant, Bank for International Settlements, Basle, Switzerland, Sept. 1980-July 1981.

Stuart M. Feder: special assignment, Bank for International Settlements, Basle, Switzerland, Jan. 1-Dec. 31, 1981.

William E. Kuhn, University of Nebraska-Lincoln: guest professor, University of Petroleum and Minerals, Dhahran, Saudi Arabia, 1980-81.

Resignations

Robert F. Allen, University of Nebraska-Lincoln: Air Force Institute of Technology, Jan. 1980.

Harald H. Boggs, American Graduate School of International Management, Aug. 8, 1980.

A. Edward Day, University of Nebraska-Lincoln: University of Louisville, Aug. 1980.

John Drake, American Graduate School of International Management, May 23, 1980.

Samuel L. Myers, Jr., University of Texas: Federal Trade Commission, Jan. 15, 1981.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution, and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Editorial Office, University of California, Los Angeles, CA 90024.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1981 issue of the *Review* will carry the seventy-eighth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1981. This announcement is an invitation to send us information for the preparation of the list. This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1981 issue the names of those who will have been awarded the doctoral degree since 1980 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

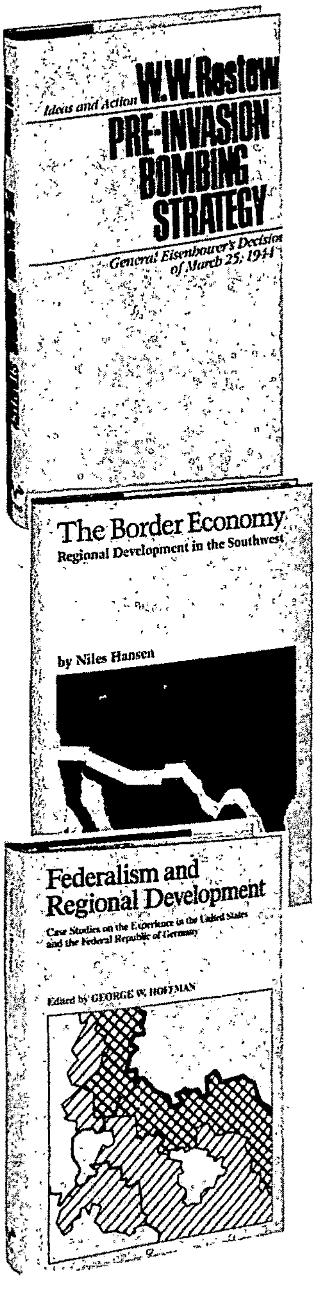
JEL Classification No. ____

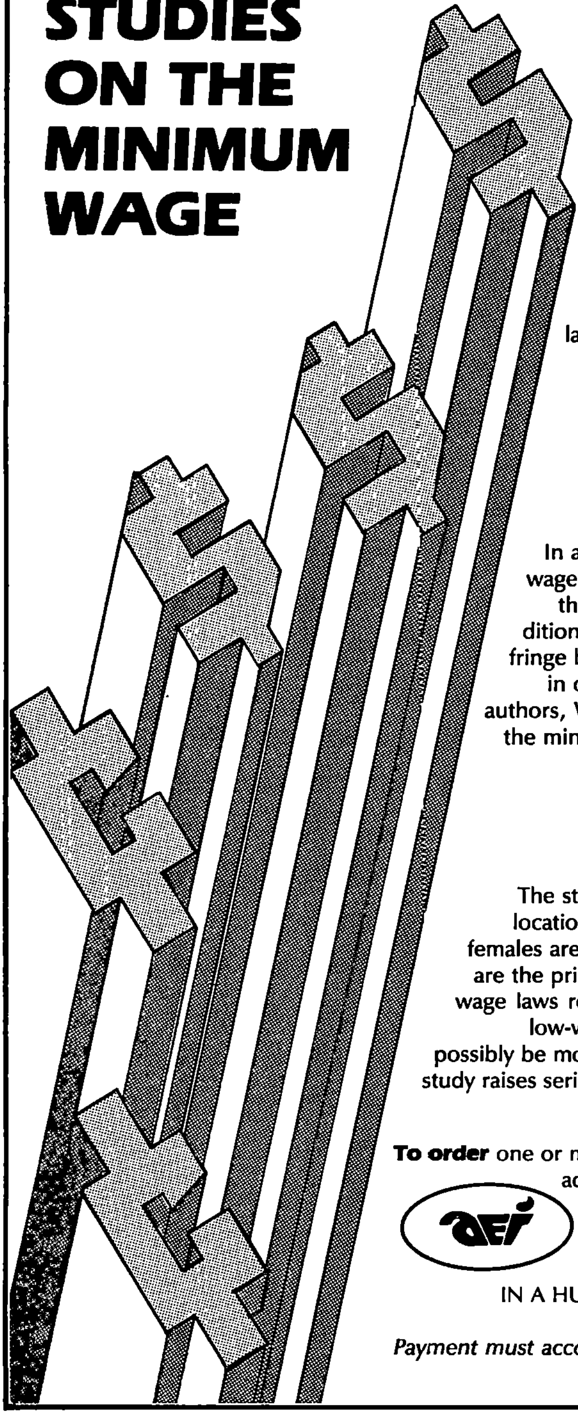
Name: LAST NAME IN CAPS: First Name, Initial ____________________

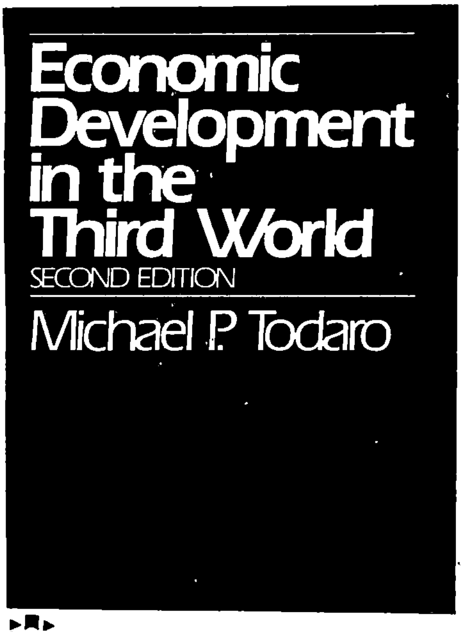
Institution Granting Degree: ____________________

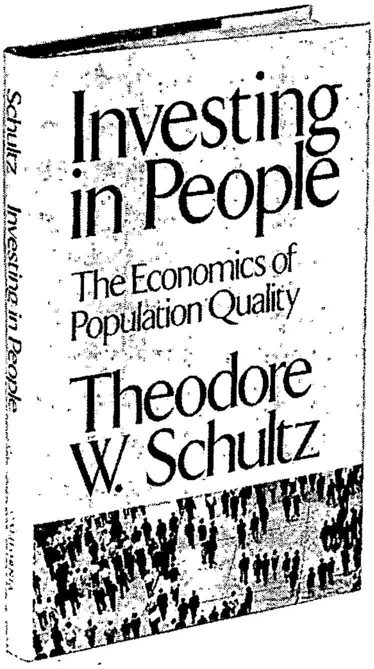
Degree Conferred (Ph.D or D.B.A.) ____ **Year** ____________________

Dissertation ____________________








Schultz Investing in People
Investing in People
The Economics of Population Quality
Theodore W. Schultz

Leading Economic Indicators for the Future....

HOW TO AVOID STRATEGIC MATERIALS SHORTAGES

Dealing with Cartels, Embargoes, and Supply Disruptions

Bohdan O. Szuprowicz

Oil—the dominant political, military, and economic tool of the 1970's—exposed the Achilles' heel of the industrialized world. In the 1980's, a whole host of strategic materials, including aluminum, titanium, cobalt, platinum, and chromium will take their place beside oil as the pivots on which the Western countries revolve. This innovative analysis of the potential for economic and political disruption vividly outlines the factors that must be continually monitored to assess the vulnerability of corporations, military establishments, and governments.

312 pp. 1981 $19.95

QUEST FOR WORLD MONETARY ORDER

The Gold-Dollar System and Its Aftermath

Milton Gilbert

A Twentieth Century Fund Study

In this candid and penetrating critique of the international money system, from Bretton Woods to the late 1970's, the late Milton Gilbert fits developments such as chronic inflation, the decline of the dollar's exchange rate, and the OPEC surpluses into an overall context.

255 pp. 1980 $21.95

PUBLIC AND PRIVATE INVESTMENTS

Socioeconomic Analysis

Henry Malcolm Steiner

This clear synthesis of project analysis and traditional microeconomic analysis bridges the gap between economic theorists and economic practitioners in government and business. It is the first thorough attempt to answer the question, "What is the connection between engineering economy, capital budgeting, and cost-benefit analysis?"

474 pp. 1980 $32.95

ECONOMETRIC ANALYSIS BY CONTROL METHODS

Gregory C. Chow

An up-to-date treatment of the techniques of stochastic control in economics. The use of control techniques for the evaluation of economic policies is now not only possible, but essential. Chow provides a rational guidebook to their effective implementation.

320 pp. April 1981 $32.95

SILVER: THE RESTLESS METAL

Roy W. Jastram

How has silver fared in periods of inflation and deflation? How are its fluctuations entwined with those of gold? What are the future prospects of "the restless metal"? These questions and many more are answered in a definitive guide to this precious metal by the author of the classic study *The Golden Constant.*

approx. 208 pp. April 1981 $26.95

SLOWTH

The Changing Economy and How You Can Successfully Cope

Martin Kupferman & Maurice D. Levi

A leading economist and a management consultant explain the real reasons behind America's slow economic growth, or *slowth,* and explain how supply-side economics can help us adjust to this disturbing new reality.

250 pp. 1980 $14.95

JOHN WILEY & SONS, Inc.
605 Third Avenue, New York, N.Y. 10158
IN CANADA: 22 Worcester Road, Rexdale,
Ontario M9W 1L1

Available at your local bookstore or use this coupon to order your books today.
For faster service call toll free: 800-526-5368.
In New Jersey, call collect: (201) 797-9809.
Order Code #2-1100. VISA, MasterCard, American Express accepted on phone orders.

Please send the books indicated for 15-DAY FREE EXAMINATION. (Restricted to the continental U.S. and Canada.)

MAIL TO: **JOHN WILEY & SONS, Inc.**
P.O. Box 092, Somerset, N.J. 08873

□ **Payment enclosed,** plus sales tax. Wiley pays normal bookrate postage/handling. We normally ship within 10 days. If shipment cannot be made within 90 days, payment will be refunded.

□ **Bill me.** □ **Bill firm or institution.**

□ Szuprowicz/STRATEGIC (1-07843-3)	**$19.95**
□ Gilbert/MONETARY (1-07998-7)	**$21.95**
□ Steiner/INVESTMENTS (1-01625-X)	**$32.95**
□ Chow/ECONOMETRIC (1-08706-8)	**$32.95**
□ Jastram/SILVER (1-03912-8)	**$26.95**
□ Kupferman/SLOWTH (1-08090-X)	**$14.95**

NAME__________

AFFILIATION__________

ADDRESS__________

CITY__________ STATE/ZIP__________

SIGN HERE__________

Prices subject to change without notice. 092 2-1100

RESEARCH GRANTS

Offered by
The Center for the Study of Futures Markets
Graduate School of Business, Columbia University

SPONSORS

Commodity Exchange, Inc. (Founding Donor)
Chicago Mercantile Exchange
J. Aron & Co., Inc.

The Center for the Study of Futures Markets offers postdoctoral and dissertation research grants to individuals interested in theoretical and empirical analyses of spot, forward, futures and commodity option markets. The Center seeks projects that adhere to the highest scholarly standards and exhibit a deep appreciation of institutional realities. Proposals are considered for funding on a continuing basis throughout the year. Typically, postdoctoral grants range from $1,000 to $10,000, and dissertation grants from $2,000 to $6,000. A recipient of a grant becomes a Research Associate of the Center and has access to the Center's data resources.

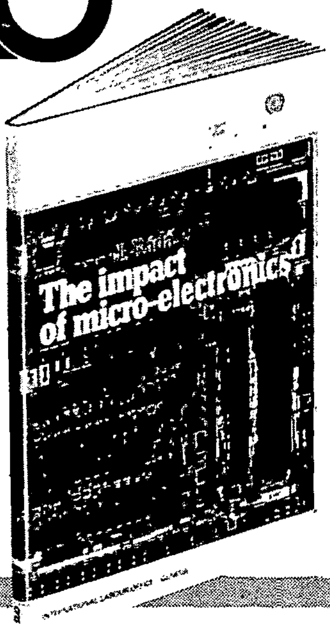
HOW TO APPLY

Proposals should be brief and include the following information:

1. A statement of the research objectives and a short literature review describing the background and importance of the study.
2. A clear and detailed description of the methodology and the expected form of the results.
3. A description of the practical implications of the project.
4. A one-to-two page abstract of the proposed project.
5. An itemized budget and time schedule illustrating the project's anticipated progress and completion.
6. A resume or *curriculum vitae* of each person who will be involved in the project, and any additional information relevant to the applicant's qualifications..
7. Graduate transcripts from university attended and letter of recommendation from thesis adviser (for dissertation grant applicants).

Please address all inquiries to:

Franklin R. Edwards, Director
Center for the Study of Futures Markets
Graduate School of Business, Uris Hall
Columbia University
New York, N.Y. 10027
(212) 280-4202


The impact of micro-electronics

AMERICAN ECONOMIC ASSOCIATION

1981 Annual Membership Rates

Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular member with rank of assistant professor or lower or annual incomes of $14,400 or less $30.00
- Regular member with rank of associate professor or annual incomes of $14,400-$24,000 .... $36.00
- Regular member with rank of full professor or annual income above $24,000 .... $42.00
- Junior member (available to registered students for three years only). Student status must be certified by your major professor or school registrar $15.00
- In Countries other than the U.S.A., Add $5.00 to cover postage.
- Family member (second membership without publications; two or more living at same address) ..... $ 6.00

Please begin my issues with:

□ **March**

□ **June** (Includes *Papers and Proceedings*)

□ **September**

□ **December** (Includes 1981 *Survey of Members*)

First Name and Initial ____ Last Name ____ Suffix ____

Address Line 1 or Attention ____

Address Line 2 ____

Address Line 3 ____

City ____ State or Country ____ Zip/Postal Code ____

PLEASE TYPE OR PRINT INFORMATION ABOVE; DO NOT EXCEED SPACES ALLOWED. DUES PAYABLE IN U.S. CURRENCY ONLY, CASHIER'S CHECK OR INTERNATIONAL MONEY ORDER PREFERRED.

Endorsed by (AEA member) ____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____

____

Authorized Signature ____

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212
U.S.A.

Begin making plans to attend the

Ninety-Fourth

Annual Meeting of The American Economic Association

(in Conjunction with Allied Social Science Associations)

to be held in

WASHINGTON, D.C.

Dec. 28-30, 1981

The Employment Center opens Sunday, December 27.

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

The 1982 meeting will be held in New York, NY, December 28-30.